CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Careless Society: Drivers of (Un)Secure Passwords

Master's thesis

Author: Bc. Vojtěch Nedvěd
Study program: Economics and Finance
Supervisor: doc. PhDr. Jozef Baruník, Ph.D.
Year of defense: 2021

Declaration of Authorship

The author hereby declares that he or she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, May 4, 2021

Bc. Vojtech Nedved

Abstract

Vulnerabilities related to poor cybersecurity are a dangerous global economic issue. This thesis aims to explain two examples of poor password management. First, why users use similar password and username and second, why they reuse their passwords, as the main drivers of this behaviour are unknown. We examined the effects of selected macroeconomic variables, gender, password length and password complexity. Additionally, this thesis suggest how to estimate sentiment in passwords using models build on Twitter posts. The results are verified on large password data, including password leaks from recent years. There are four main findings. First, a higher cybersecurity index and diversity of a password seem to be related to the lower similarity between a username and a password. Second, it seems that there are structural differences between countries and languages. Third, the sentiment seems to be a significant determinant too. Fourth, password reuse seems to be positively affected by the cybersecurity level. The thesis contributes to the study of password management. It proposes how to model the relationship, derive the data, split the passwords into words, model the sentiment of passwords, what variables might be used and how the results might contribute to better password policies.

JEL Classification	C01 C55 D90 D91 F52 J18
Keywords	password policy, cybersecurity, behavioral economics, digital economy
Title	Careless Society: Drivers of (Un)Secure Passwords

Abstrakt

Hrozby vycházející z nedostatečné kyberbezpečnosti mohou být nebezpečným fenoménem ohrožujícím ekonomické zájmy. Cílem této práce je popsat dva příklady rizikového chování uživatelů: 1. Proč se užívají podobná uživatelská jména a hesla. 2. Z jakého důvodu uživatelé svá hesla recyklují. Zaměřili jsme se na vliv několika makroekonomických proměnných, pohlaví, délky, komplexity hesla a sentimentu. Navíc v této práci uvádíme i příklad jak sentiment v heslech detekovat. Vztahy mezi proměnnými byly ověřeny na základě velkého vzorku dat hesel posledních let. Sentiment byl určen pomocí vytvořených modelů na základě příspěvků z Twitteru. Výsledky se dají shrnout do čtyř hlavních bodů: 1. Vyšší kyberbezpečnost je spojena s nižší podobností hesel a uživatelských jmen. 2. Výsledky naznačují systematické rozdíly podobností napříč zeměmi a jazykovými skupinami. 3. Sentiment hesla má také významný vliv na podobnost hesel a uživatelských jmen. 4. Recyklace hesel se zdá být pozitivně ovlivněna kyberbezpečností, vedoucí k nižší míře recyklace. Tato práce přispívá k výzkumu používání hesel. Ukazuje, jak by se dané vztahy mohli modelovat, jak extrahovat informace z dat, rozdělit heslo do slov. Jak modelovat sentiment hesel, jaké proměnné by mohli ovlivňovat management hesel a jak by výsledky mohly přispět k lepší kyberbezpečnosti pomocí silnějších hesel.

Klasifikace JEL	C01 C55 D90 D91 F52 J18
Klíčová slova	tvorba hesel, kybernetická bezpečnost, behaviorální ekonomie, digitální ekonomie

Název práce	Nepoučitelní uživatelé: příčiny (ne)bezpečných hesel

Acknowledgments

I would like to express my gratitude to my thesis supervisor doc. PhDr. Jozef Baruník, Ph.D.for his valuable and insightful advice in preparation of this thesis and, furthermore, to doc. Mgr. Barbora Vidová Hladká, Ph.D. and RNDr. Milan Straka, Ph.D. for their priceless advice on the Natural Language Processing part. Finally, yet importantly, I would like to express gratitude to my family for their continuous support during my studies.

Typeset in LaTeXusing the IES Thesis Template.

Bibliographic Record

Nedved, Vojtech: *Careless Society: Drivers of (Un)Secure Passwords.* Master's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2021, pages 284. Advisor: doc. PhDr. Jozef Baruník, Ph.D.

Contents

List of Tables

List of Figures

Acronyms

AI Artificial Intelligence

BMA Bayesian Model Averaging

BNC Birtish National Corpus

COCA Corpus of Contemporary American English

NLP Natural Language Processing

PI Personal Information

POI part-of-speech

PPSim Password-Password similarity

PUSim Password-Username similarity

TLD Top Level Domain

UD Universal Dependencies

WBA Word Break Algorithm

Master's Thesis Proposal

Author	Bc. Vojtěch Nedvěd
Supervisor	doc. PhDr. Jozef Baruník, Ph.D.
Proposed topic	Careless Society: Drivers of (Un)Secure Passwords

Motivation Authentication plays an important role in protecting online accounts such as a bank account, social network account or an email account. The current authentication technology consist most frequently on (a) a password, which is a sequence of letters, numbers and special characters and (b) a biometric solutions such as retina scan, fingerprint scan or 3D face recognition.

When the authentication is solely based on a sequence of letters and numbers, users face two key vulnerabilities. Users rather choose simple weak passwords (i.e. short strings frequently composed by only letters) in opposition to complex and long passwords (including numbers and special characters) as they are hard to remember. Moreover, if they choose a complex and secure password, they are likely to forget them. Password managers help users with both issues. They allow them to use complex secure passwords on one hand, and on the other, they do not rely on their memory to remember them. Unfortunately, even this solution might imply a strong vulnerability. All these passwords are backed up by only one master password, that allows users to access all their password. That means, if someone could hack or guess the master password, all user's passwords would be leaked and the hacker might cause irreversible damage by stealing money from a bank account or misuse the user's identity.

The biometric approach of authentication, recently rising in popularity, helps to avoid these vulnerabilities caused by a weak password. Fingerprints or retina scans are unique in the population and are decently hard to break. Nevertheless, they are frequently backed up by a password that is used when the biometric scans fail. Thus, those passwords that serve as a backup for biometric authentication might present a significant threat if the user is not responsible enough and chooses a simple password.

To sum it up, even with password managers and biometric authentication, passwords still play an important role in cybersecurity and in the case of password managers, their importance is even higher than before as one password is a key to all of them. We need to study our behaviour related to the password management as all potential vulnerabilities (i.e. high rate of reuse of passwords, relation between username and password or frequent simple words included in the password) might be exploited by attackers.

Researchers usually try to assess the strength of the password by statistics evaluation (i.e. entropy) or using dictionaries of languages that assess whether a particular password contains common words or not. Those passwords are vulnerable to so-called dictionary attacks. To improve the security of users, providers of services that use passwords require a set of rules to improve the passwords (i.e. length, usage of upper and lower case, special symbols). Moreover, companies usually impose several rules that improve security such as non-repetitiveness of passwords, regular changes of passwords or non-similarity of historical and new passwords. The measurement of the success of those policies is rather dubious. Data on passwords are very sensitive and IT departments usually do not even store passwords but their hashed version which means using a conventional computing power no-one can reverse the hashing process and get the original password.

We know that people tend to be irresponsible and frequently reuse their password. They reuse it among one web service upon request to change their password given the policy rules or they reuse them among different web pages. Would it be possible to study this behaviour? Are we able to identify some drivers that influence this careless behaviour? Are there differences among nations and languages? Is it affected by literacy rate or even democracy level? Secondly, what is the relationship between the username and password? People tend to be lazy and we expect a strong and frequent similarity between the username and password. Does it differ among nations, sex, age or literacy? And thirdly, we are curious about the sentiment in the password. There are nations that tend to be pessimistic. How is it reflected in passwords? Are countries where the negative sentiment prevail in password? Or are there countries that tend to be neutral? The sentiment is another vulnerability that might be used as a weapon by hackers and we would like to study this perspective as there is not much known about it.

To our best knowledge, an econometric assessment of password management behaviour was not done before. Subsequently, we have found small research on sentiment analysis in Chinese passwords but without inference with econometrics.

The motivation behind this topic is to try to estimate relationships in a field that is poorly studied from the causal point of view and is of extreme importance in regard of our online security. Additionally, we will try to blend modified Sentiment Analysis with econometrics. The thesis will be challenging in terms of variable and model selection, data preparation for the Sentiment Analysis and the data processing itself as we are dealing with a decently large dataset.

The goal of this thesis is to identify key macroeconomic and other variables related to level of password recyclation. We will try to identify factors that might have an effect on this behaviour such as gender, age, nationality, language, literacy rates or democracy. Furthermore, we are curious about the relationship between the username and the passwords and the drivers that might affect it negatively. Lastly, we will study the sentiment hidden in password and its effect on reusing the password. We will refer to this sentiment as positive or negative connotations as for a standard Natural Language Processing (NLP) Sentiment Analysis we would need rather sentences to capture the context. We will deal with words or potentially couple of words. This task will be performed on selected groups of languages: the Czech Republic, Spain, selected countries from Latin America and selected English speaking countries such as the United Kingdom.

Hypotheses Following the previous introduction to the topic, we will focus on three key areas. In the first part, we will examine the factors that might be affecting the reuse of passwords. Initial variables to be considered are literacy rates, digital education, number of devices per user, number of computers attacks per capita, the complexity of language, GDP, development index, estimated age of the user, estimated gender and provider of the email. However, we will also include variables such as personal information or keyboard patterns.

Let PS be a level of similarity among two passwords from a single user. The hypothesized model has following form:

$PS \sim education + literacy + freedom + language + country +$
$general\ development\ of\ the\ country + password\ features + user + cyber\ security$

Note: password features are, for example, length, use of upper/lowercase/special characters, use of foreign language

In the second part, we will study the similarity between username and passwords. We will estimate the similarity using Levenhstein distance and test several variables that we believe might have some impact. Variables considered for testing are the same to the hypothesis one as the dependent variables are similar in nature and we are exploring what might be significant without previous research in this field.

Let US be the similarity between the username and the password. The hypothesised model has following form:

$US \sim education + literacy + freedom + language + country +$
$general\ development\ of\ the\ country + password\ features + user + cyber\ security$

In the third part, we will try to assess the sentiment of the passwords. The first key potential finding is to confirm the significant presence of some sentiment in the selected languages. That will be further extended to testing whether pessimist nations tend to use negative vibes in their passwords.

Let S $\in$ negative, neutral, positive be the estimated sentiment of the password. The hypothesised model explaining the occurrence of the sentiment (binary variable) has following form:

$Sentiment \sim country + language + education + democracy + user$

The hypothesised models will be verified by econometric tools. Because of the lack of previous causal examination of the variables, models are subject to change.

Methodology Password data are very sensitive and there are very few (if none) official data set. However, there are a number of data leaks from several providers (e.g. linkedin.com, seznam.cz, google.com, facebook.com and many more). We have acquired a decently large dataset that appeared on web around 2015. This breach contains famous leaks such as RockYou or Linked in database of passwords. The same breach was analysed by Li et al. (2019). The data contain over 1 400 million emails and their corresponding passwords across the globe.

The first step in the analysis will be to transform and clean the data. Records will be parsed using regex into the following parts: a) username b) provider c) domain (first, second and country level) and d) password. After the cleaning, data will be moved to a high-performance database for efficient computing, possibly PySpark technology. For the Sentiment Analysis part, we will need to identify meaningful words from the passwords what will be done by some algorithm comparing strings with the vocabulary of given language. In this part we will face significant computational limits.

For the similarity, we will apply Levenhstein distance to determine how likely two passwords are. The Levenhstein distance will be also used for assessing the similarity between username and password. However, in this case we will also consider Longest Common Sequence (LCS) to identify the longest string possible that appears both in the username and password.

Microeconomic factors such as gender or age will be derived from the parsed information. Gender will be assessed (with some inaccuracy) based on dictionary comparison with names of given languages and the username. Age will be guessed (with some inaccuracy) based on numbers appearing in the username (and potentially passwords as well).

The sentiment of the passwords will be identified by writing an algorithm for a word identification in a string using dictionaries of selected languages. It will be taken into consideration whether to use a conventional sentiment analysis or use a modified one which we called vibes that indicate the positive or negative context of a given word.

Hypothesis will be tested using linear regression estimators and logistic regression. We will be specifically careful about random sampling. Bootstrapping is one of the options we will implement.

Expected Contribution Very few research is done on password use behavior from an econometric point of view. A number of reports containing descriptive statistics could be found on the internet (delivered both by private security companies and research institutions). However, we have not found any detailed research investigating why this behavior happens. Thus, this thesis will be one of very few research focusing on determining factors that affect the above-mentioned variables such as the reuse of passwords, username and password similarity and sentiment occurrence. It will be analyzed using an extensive dataset containing over 1 400 millions of passwords. Additionally, we will blend Natural Language Processing (Sentiment Analysis) with econometrics and additional smart algorithms to get the maximum information out of usernames and passwords. Last, but not least, we will cover the Czech language, which is usually not popular for this kind of study due to the small population, and a group of Spanish countries and some English speaking countries. The ultimate contribution will be the proposed models to describe people's behavior behind password management.

Outline

1. Introduction
2. Literature review
   (a) Online security
   (b) Passwords in general
   (c) Available descriptive statistics
   (d) Related research
3. Methodology and data
   (a) Data
   (b) Approach for testing H1
   (c) Approach for testing H2
   (d) Approach for testing H3
   (e) Modified semantics (Nature Language Processing)
   (f) Approach for testing H4

   (g) (Under consideration) Approach for testing H4 (drivers of the strength)

4. Empirical model
5. Discussion of results
6. Conclusion

Core bibliography

LI, Yue, Haining a Kun SUN. A study of personal information in human-chosen passwords and its security implications. IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications. 2016, 2016, , 1-9. DOI: 10.1109/INFOCOM.2016.7524583. ISBN 978-1-4673-9953-1. Dostupné také z: http://ieeexplore.ieee.org/document/7524583/

Houshmand, Shiva & Aggarwal, Sudhir. (2017). Using Personal Information in Targeted Grammar-Based Probabilistic Password Attacks. 285-303. 10.1007/978-3-319-67208-3_16

Bulbulia, Z, and M Maharaj. "Factors That Influence Young Adults' Online Security Awareness in the Durban in South Africa." Journal of Information Warfare, vol. 12, no. 1, 2013, pp. 83–96. JSTOR, www.jstor.org/stable/26487001.

Bulbulia, Z, and M Maharaj. "Factors That Influence Young Adults' Online Security Awareness in the Durban in South Africa." Journal of Information Warfare, vol. 12, no. 1, 2013, pp. 83–96. JSTOR, www.jstor.org/stable/26487001.

Yu, Xiaoying & Liao, Qi. (2019). Understanding user passwords through password prefix and postfix (P3) graph analysis and visualization. International Journal of Information Security. 10.1007/s10207-019-00432-3.

BONNEAU, Joseph. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. 2012 IEEE Symposium on Security and Privacy [online]. IEEE, 2012, 538-552 [cit. 2019-07-12]. DOI: 10.1109/SP.2012.49. ISBN 978-1-4673-1244-8.

Li, Zhiyong & Li, Tao Zhu, Fangdong. (2019). An Online Password Guessing Method Based on Big Data. 59-62. 10.1145/3325773.3325779.

Chapter 1

Introduction

There is a lack of understanding of why people choose a poor password. It has been shown multiple times that user's passwords are often vulnerable (Šolić *et al.* 2015; Weber *et al.* 2008; Ur *et al.* 2019). Surveys are trying to describe people's attitude to passwords (Helkala & Bakås 2014; Haque *et al.* 2014) but they are frequently based on descriptive statistics and they do not overcome the issue that users might be reluctant to share their actual behaviour in a survey (Brenner & DeLamater 2016). However, the causal reasons why users use poor passwords are not completely clear.

The understanding of peoples behaviour is crucial for designing effective password policies. The standard recommendation of having long passwords composed of letters, numbers and symbols does not ensure the password will not be derived from the username or formed by frequently used words. Users might be prone to follow different bad practices, and providers might tailor password policies aiming to eliminate this behaviour.

In today's society, the secured digital product and services are one of the vital aspects for maintaining trust in the system. People use e-mails for private communication, access bank accounts and communicate with bureaus through these electronic systems. Cybersecurity is linked with users, companies and governments, playing an essential role in the modern economy (Moore 2010; Gour 2014).

This thesis aims to explain two examples of poor password management. First, it tries to identify drivers of the similarity of a username and a password. Second, attempts to find variables affecting the reuse of passwords. Both are poor practices negatively affecting the security of the accounts.

The Password-Username similarity and the Password-Password similarity were explained by a set of macroeconomic variables (i.e., Literacy, Internet coverage, Democracy, Mobile phone usage and Cybersecurity). Additionally, it was estimated the effect of password length, gender and polarity (i.e., positive or negative connotations) of a password on the Password-Username similarity.

The password data were composed by a set of famous data leaks such as LinkedIn leak (Kontaxis *et al.* 2013), RockYou leak (Yan & Chen 2018) or Yahoo! leak (Zhang *et al.* 2019). Over 1.4 billion observations were processed and sampled to 2.5 million observations to match the available computational power. A couple of the variables had to be derived from the password data. The gender was estimated using a list of scraped names in dozens of languages. The diversity of a password indicated how many character groups are present in the password (i.e., letters, numbers and symbols).

Regarding the polarity, passwords were broken into words using a custom-built word break algorithm and statistical language models. These words were labelled as positive, neutral or negative using a logistic regression trained on 700 thousand Twitter posts for nine languages.

The effects of these variables were estimated using the Generalised Ordered Logistic Regression. The results suggest that the cybersecurity index is an important variable. Better cybersecurity seems to be associated with higher dissimilarity between a username and a password. Similarly, higher cybersecurity seems to be associated with lower password reuse. Furthermore, the password diversity seems to be related to the higher dissimilarity between usernames and passwords.

The results also suggest that the polarity affect the Password-Username similarity. It seems that the presence of a polarity in the password is associated with a higher similarity between usernames and passwords.

As expected, the results also suggest that there are structural differences among languages and countries. For example, Czech users have, overall, more similar passwords and usernames than German users. Similarly, users from the Slavic language group have more similar passwords and usernames than users from the Germanic language group.

These results might be valuable for designing more effective and targeted password policies. Providers, such as Google, might set up password policies based on the level of cybersecurity in a given country. If there is low cybersecurity, it will make sense to remind users not to derive their password from a username.

Similarly, providers should continue to encourage people to use various characters in their passwords. Results suggest that diverse passwords seem to be more secure.

The Internet coverage, Literacy rate and Democracy level also seem to be associated with the Password-Username similarity. After additional research, these variables might be used for tailoring the password policies similar to the cybersecurity index.

There seems to be differences among countries in terms of their attitude to password management. Further investigation might reveal the most vulnerable countries, how they differ from the most responsible countries and how the findings could be leveraged for improving the password policies.

The thesis contributes to the study of human behaviour when managing passwords. It was proposed how to model such a relationship, derive the data, split the passwords into words, model the polarity of passwords, what variables might be used, and, most importantly, that these findings could improve the password policies. Hence, improving the security of digital services rooted in today's economy.

The thesis is structured as follows: Introduction motivates the study of the problems, Literature review gives an overview of existing research, Methodology describes the statistical approach, Results presents the main findings, Robustness check examines the robustness of the results, Discussion shows what might be done differently and last, Conclusion summarises the problem and the findings.

Chapter 2

Literature Review

This chapter gives an overview on the existing research on passwords management. First, we discuss the role of passwords among today's authentication choices. Second, we briefly describe how passwords managers help users. Third, we show results of current research related to passwords. Fourth, we summarise the findings related to password reuse. Fifth, we briefly touch sentiment in text. Sixth, we describe current understanding of language structure of passwords. Last, we present hypothesised models to be investigated.

2.1 Authentication systems

Passwords are no more the only mean of authentication. Biometric solutions developed years ago are becoming popular. Most modern smartphones feature fingerprint sensor or face scan. In that case, one might question the importance of passwords. Nevertheless, passwords might have several benefits and might not be replaced anytime soon.

Ten years ago, Bonneau *et al.* (2012) presented a comparison of present authentication systems. They claimed that despite having complex and possibly secure biometric solutions, passwords would be difficult to replace. They argued that even though their colleagues were sceptical about the future of password use, passwords would not have been replaced any time soon due to many real-world constraints.

Today, nearly ten years after the analysis, we can confirm their expectations. Even though many people use fingerprint sensors and face scans, passwords still dominate the authentication solution, especially on computers.

The range of currently available solutions became wide. Users can choose to protect their accounts by a password, fingerprint scan (Ogbanufe & Kim 2018), retina scan (Mazumdar 2018), face scan (Fathy *et al.* 2015) or one can use the palm to verify the identity (Shinzaki 2020). Users might even use the dynamics of keystrokes on the keyboard for authentication Tey *et al.* (2013).

Despite the numerous options, fingerprint or face scans on smartphones are backed up by a standard password (both on an Android and Apple device). The reason is apparent. Despite relatively high accuracy of verification (Mathur *et al.* 2016), passwords serve as a backup key to the service. If the user would be injured, the password serves as a non-dependant way of verification. Moreover, one should not forget that stolen password can be easily replaced.

There are two main reasons why users data might be stolen. First, a hacker attacks the system and steals password. Second, users use weak passwords that are easy to guess.

Several password leaks happened throughout the history (Heen & Neumann 2017), and it is not in the power of users to prevent their passwords to be stolen. It is the responsibility of the provider to secure the server where passwords are stored. In 2009, RockYou web service was hacked and millions of unencrypted passwords leaked, and they can be found online[1]. On the other hand, what a user can do is to create a strong password so that it would be difficult for hackers to crack it (Shay *et al.* 2016).

Despite the popularity of biometric solutions, some researchers believe that passwords will continue to play an essential role in the authentication until several constraints are solved. First, Pareto-improving verification is developed, and second, users are motivated and convinced to abandon textual passwords (Bošnjak & Brumen 2019).

2.2 Password management

The number of online services used by people is increasing steadily. A few years ago, users had an email account, an account for their popular e-shop or a bank account. Nowadays, we log in to social sites (e.g., Facebook, Twitter, Instagram), pay using online bank accounts, use several email addresses, shop in multiple e-shops and sign in to work accounts. The number of accounts one operates increased dramatically. Stobert & Biddle (2018) report that people operated with 9 to 51 accounts. In a slightly older study, Florencio & Herley

[1]Description available at https://www.kaggle.com/wjburns/common-password-list-rockyoutxt

(2007) estimated that an average user had around 25 accounts. That naturally implies pressure on both password creation behaviour and their management. A large number of passwords might be harder to remember.

Passwords managers help users to store their passwords securely. Furthermore, their advantage is that they make it feasible to use hard to remember passwords (Alkaldi & Renaud 2016). While there might be a perception that these managers are highly secure, several studies deny this claim (Chiasson *et al.* 2006; Li *et al.* 2014; Belenko & Sklyarov 2012).

Empirical studies show that password managers positively affect password quality. However, it does not imply that the user would use stronger passwords (Lyastani *et al.* 2018). Regardless of the perception of the security of managers, the access to the password manager is frequently arranged by one master password.

The idea is to have a single hard-to-break password that keeps the rest of the passwords secure. Naturally, these single passwords might present a vulnerability to the account. If a user would not choose a strong password, then the overall security of the portfolio of his passwords would decrease. One password with a below-average difficulty to crack would disclose all the user's passwords. Thus, it would be wise to have a very strong master password.

2.3 Password analysis

Given the ongoing research on password strength and the potential vulnerabilities, researchers aim to analyse passwords and recommend best practice on password security or develop a better cracking algorithm presenting a thread to the users.

Shen *et al.* (2016) studied over 6 million passwords to present an in-depth analysis of user practice in real-life passwords. They focused on password length, password composition and password selection. Their conclusions are following. First, the average password is at least 12% longer than reported in older studies. Second, they found an increase in the proportion of passwords consisting only of numbers. Third, they found a significant increase in passwords containing the corresponding username or well known weak passwords.

They concluded that there is a shift in password habits over time and that the results differ from the survey-based analysis of user customs.

Rao *et al.* (2013) studied the relationship of grammatical structures with password strength. They showed that the search space for guessing a password might be decreased by up to 50% if the grammatical structure is considered. Furthermore, under the assumption that longer password tends to be more secure, they concluded that the strength of long passwords does not increase uniformly with length. Finally, they presented a grammar-aware cracking algorithm claiming that it cracked 10% of passwords more than state-of-the-art password crackers. This study points out the importance of language and its grammar for password modelling.

There is further evidence of password variation over time. von Zezschwitz *et al.* (2013) investigated the evolutionary change of user-selected passwords. Data were obtained during surveys with users aiming to examine password reuse, password changes, and factors influencing the password life cycle. The authors concluded that the latest passwords are substantially longer than the very first ones. Furthermore, even though users knew how to construct a secure password, they refused to follow good practice and used alarmingly weak passwords for most services.

Furthermore, the authors found similarities among passwords of a single user. They report frequent modification of the first password as a root for next-generation passwords. These passwords were accepted by the systems despite password policies and meters.

Researchers also focus on the identification of personal information employed in passwords. Bulbulia & Maharaj (2013) published a paper aiming to explain what factors influence online security. They took a pool of young people from Durban, South Africa and conducted a statistical analysis using the approach proposed by Maddux & Rogers (1983).

Data were obtained through an online questionnaire sent to the university students. The most significant findings were that race, gender, and employment status have a strong relationship with online security awareness. The influence of gender on online security is also confirmed by McGill & Thompson (2018), concluding that the overall security level for females is lower than for males. On the other hand, the authors claim that females tend to have a higher level of awareness of security threats than males.

Petrie & Merdenyan (2016) also studied possible differences caused by gender. Having a sample of 202 women and men from three different countries (China, Turkey and the UK), they did not find strong support for cultural differences in password management and security awareness. However, they suggest that cultural background and gender should be considered when explaining users' password choices.

Li *et al.* (2016) brings another evidence on differences among males and females. With a paper aiming to take advantage of the personal information incorporated in a password for improving the PCFG guessing method, they conclude that the length of a password among males and females does not differ on average.

Nevertheless, the occurrence of personal information is estimated to be by six percentage points higher for males than for females. That would indicate a more responsible attitude of females than of males. This finding is in contrast with the conclusions of McGill & Thompson (2018) claiming that security awareness is lower for females in comparison with males.

The lower security of male's passwords is also suggested by Bonneau (2012). The authors also suggest that password strength is positively correlated with age, implying better security for older users. The results also suggest differences in security among cultures. According to the study, Indonesian-speaking users were among the less responsible, while German and Korean-speaking users exhibited the best password's strength in the sample.

2.4 Reuse of passwords

One of the unpleasant habits users poses is the use of a password across multiple platforms. They might generate one password and use it frequently. That might present a threat to the security of their accounts. If the user were subject to a password data leak, the hacker would take over the original account and breach into linked accounts. A few studies aim to understand this behaviour and describe what is related to the reuse of passwords.

Researchers focus on the characteristics of the poor behaviour but might omit the study of why is it happening. Because of the frequent use of interviews to assess people's behaviour, they work with a relatively small number of observations and artificially created data during the interviews (Das *et al.* 2014; Haque *et al.* 2013; Komanduri *et al.* 2011a; Notoatmodjo & Thomborson 2009). One might be sceptical about the reliability of produced data during the interviews. It has been shown that users responses in surveys might not reflect their actual behaviour (Brenner & DeLamater 2016).

Wash *et al.* (2016) studied how frequently are passwords reused across web sites. They combined self-reported survey responses with data gathered from online websites over the course of six weeks. One hundred thirty-four participants were included in the study. They estimated that users tend to reuse a single password 1.7-3.4 times across different websites. Additionally, they suggest that people mostly reuse complex passwords while keeping short and straightforward passwords unique per page.

Trying to understand self-reporting behaviour, they also focused on comparing self-reported measures with data gathered from the websites. Interestingly, they found out that users' awareness about password reuse risks is high. However, users do not translate the awareness into practice. Despite the known risks, they follow bad habits using the same password on multiple pages.

Their results indicate that to solve the inconvenience of having multiple passwords, people tend to create a single strong password including upper and lower cases, numbers and special characters. However, if they do so, they are more likely to use the password for multiple accounts. On the other hand, users with short, weak passwords tend to manage a higher number of them. Contrary to these results, authors suggest that the reuse of strong versus weak password works in the opposite way (Stobert & Biddle 2018).

The findings might be expected. Users can come up with a decent password. However, they tend to reuse it frequently and users seem to be aware of their security. Nevertheless, the translation into practice is dubious. The strength of this study is the detailed data researchers had about the users. They captured the behaviour every time the user attempted to access a website. The average user in the study visited 5 613 pages which imply 118 web pages per day. Not all visits are connected to password use. Thus, the frequency of password fillings ranges from only 22 password entries to a maximum of 1 474 entries. That is a very good granularity of data about the behaviour of the participants.

One of the weaknesses of the study is the selected population. The survey was conducted on university students. Thus, we would expect that the results do not generalise to a broader population as it was shown that older people tend to use stronger passwords (Bonneau 2012).

Hanamsagar *et al.* (2016) conducted an in-depth study on password reuse, including 50 participants in the survey. They aimed to study the semantics structure, strength and reuse of the passwords. They found that an average password is weak and is connected to more than four sites. Surprisingly, for essential web sites (e.g., banking), the passwords tend to be only 1-2 longer and ten times stronger than for casual sites. 84% of users reuse passwords between unimportant and important pages, increasing the vulnerability of essential accounts.

Regarding the similarity among passwords, 98% users reuse the same password, and the remaining 2% apply only minor changes. They identified the misconceptions about the risk and preference on memorability over security as the main reasons for frequent password reuse.

In contrary to their results, Das *et al.* (2014) estimated that 43-51% of users keep an identical password for multiple accounts. Users included in the study had an extension to their browsers to collect passwords. In order to ensure data privacy, passwords were anonymised on the subject's computers, and researchers were provided only with the anonymous structure form. They used semantic segmentation, and Part-Of-Speech (POS) tagging originally developed by Veras *et al.* (2014). They took a password and preprocessed the string following KoreLogic to crack passwords [2].

They demonstrated it in the following example. Having a pair of passwords "john352@" and "john222", the splitting method would result in (proper-name)(3- digit-number)(special-char) and (proper-name)(3-digit-number). This approach suggests one of the possible manners how one might split a password into meaningful words. The authors used this technique solely in order to encrypt the user's password.

The mentioned studies focusing on the reuse of passwords deal mainly with reusing the same password. A second option is to measure reuse, including a modification of the password (Wang *et al.* 2018). A user would take an existing password, slightly modify the string and use it in a different service. This behaviour is slightly less insecure than the exact reuse. However, still present a serious threat to the security as a hacker might derive characteristics of other passwords by having one of them.

There are also studies based on more significant amounts of observations. Wang *et al.* (2018) examined a decently large data set containing 28.8 million users with 61.5 million passwords covering 107 services during eight years.

As expected, they found a high rate of reuse of passwords. They estimate that 52% of the users involved in the study exhibit some sort of password reuse. They identified email accounts and shopping websites as places with the highest reuse rate. Surprisingly, a number of users reused the password even though they knew their credentials leaked in one of the reported password leaks.

The information obtained in the study was used to improve a guessing algorithm, leaving the identification of drivers of reuse behind. However, they implement an algorithm to detect how two passwords differ. The considered modifications are following:

[2]https://contest-2010.korelogic.com/rules.html

1. Two passwords are identical
2. One is a substring of the second
3. Capitalisation was applied on the first one to obtain the second one
4. Is the second one a reversal of the first one?
5. Is some sequential algorithm applied to generate the second one?
6. A combination of above
7. Cannot find a rule

In 34% of the cases, the two considered passwords were identical. The second highest pattern was a substring followed by the capitalisation modification. Nevertheless, in 46% of the cases, no rule was identified.

Based on these observations, authors created a reuse rate and a modification rate. They concluded that both rates are increasing as users use more accounts with passwords. Thus, security estimations that do not take into account the modification will severely underestimate the security risks. The information obtained from a thorough study of the passwords was afterwards used to guess the password. Authors showed that with a relatively small sample size they can achieve a similar performance to conventional password cracking algorithm.

A few researchers also focused both on password reuse and modification. Nevertheless, frequently the only purpose is to describe the level of reuse among the sample. In a smaller number of papers, researchers use identified password features to explain the reuse and afterwards, they used the patterns for improving password cracking. Furthermore, most of the studies are based on interview data, and fewer studies use an empirical approach on large amounts of data. Up to our best knowledge, there is no study aiming to identify sociological factors affecting this phenomenon. Thus, it is unclear why is it happening, what affects the modification and exact reuse and how one might prevent it.

2.5 Sentiment in text

There is extensive research on studying sentiment in a text. Mäntylä *et al.* (2018) recently evaluated research in the sentiment analysis and identified current challenges. They analysed 6 996 papers from Scopus dedicated to the topic of sentiment analysis and its detection. They found the roots of the domain in public opinion analysis at the beginning of the 20th century, with computational linguistics community involvement in the 1990s. They claim that 99% of the papers have been published after 2004. According to the authors, analysis has shifted from analysis of product reviews to social media texts from Twitter and Facebook, stock market predictions, elections predictions, disasters forecasting and applications in healthcare.

Classic sentiment analysis frequently requires a coherent text to understand the context and estimate the sentiment. Nowadays, attention is also put on challenging areas of sentiment analysis, including short texts. Davidov *et al.* (2010) studied the sentiment of Twitter data by enhancing the common estimation by examining hashtags and emoticons. They presented a classification framework with 50 Twitter tags and 15 smileys as sentiment labels.

Taboada *et al.* (2011) studied sentiment under the constraint of limited text too. They took a lexicon-based approach to extract sentiment from a text. The paper's core is the Semantic Orientation CALculator (SO-CAL), which can work with dictionaries with annotated words, indicating their polarity and strength. They demonstrated that the technique is consistent across domains and even on unseen data making the approach highly robust. Furthermore, they present a methodology on how to construct a good dictionary for building the SO-CAL.

Yi *et al.* (2003) developed a Sentiment Analyser capable of extracting a sentiment about a subject from an online text. As opposed to traditional techniques (Mr. S. M. Vohra 2012), authors did not detect sentiment on the whole text but instead identified the subject of the text and evaluated all references to that subject.

Social media, especially Twitter, are exceptionally popular for sentiment analysis. Estimation of the relationship between tweets' sentiment and stock prices (Bakshi *et al.* 2016; Mittal & Goel 2012), prediction of election results given the sentiment of tweets (Tumasjan *et al.* 2010), identification of a public opinion on a brand given customer tweets (Ghiassi *et al.* 2013) and studies were aiming to improve the classification of the sentiment (Kontopoulos *et al.* 2013; Saif *et al.* 2012; Giachanou & Crestani 2016; Thelwall *et al.* 2011; Jianqiang *et al.* 2018).

The understanding of a language might require an effort even for humans. For example, when people use irony and sarcasm. Filatova (2012) worked with Amazon reviews data to study how to work with irony and sarcasm in a text.

They presented a corpus generation experiment where they collected product reviews from Amazon and then performed a qualitative and quantitative analysis of the resulting corpus. They created a corpus that can be used for sarcasm detection using long texts or short sentences. This could serve social media analytics or help chatbots interact appropriately with humans when they do not mean what they write.

Sentiment analysis is still popular even though it began decades ago. There is room for improvements of the models, and researchers investigate where the application might be helpful—for example, elections prediction and stock market forecasting.

2.6 Language understanding of password

Psychologists were one of the first researchers investigating the sentiment of passwords. Brown *et al.* (2004) used a survey approach to identify personal information included in the passwords. They found that names are the most frequent class of words in the collected sample, followed by names of family, relatives and friends.

Riddle *et al.* (1989) also used university students to assess the common words in passwords. They arrived at slightly different conclusions. They found that birth dates, personal names, nicknames and celebrity names are the most frequent elements of passwords.

The sample size both research groups used was decent. In the case of Riddle *et al.* (1989) 6 226 subjects were included. One of the concerns might be the sample. One could argue that the university environment is specific due to the above-average educated population and age range (young students).

Pilar *et al.* (2012) run a study investigating the effect of age and education using ANOVA. They confirmed a positive effect of education on password strength. However, they failed to find a significant relationship with age. A report indicating potential differences among ages was found, but without a strong credibility[3].

The occurrence of a name in the password is also supported by Bonneau & Shutova (2012). They investigated how the occurrence of common categories of words (e.g., musicians, albums, names, books, brand names) influence the strength of a password. For example, a city or a state in the USA appeared in 0.8% of the cases. They emphasise that in 4% of cases, a person's names was

[3]https://digitalguardian.com/blog/uncovering-password-habits-are-users-password-security-habits-improving-infographic

found as a part of the string. That is an alarming figure, as the guess-ability is exceptionally high.

The structure of passwords is often studied with the goal of developing a model for efficient password cracking (Malone & Maher 2012; Ma *et al.* 2014). There is a lack of research on the semantic or lexical content of passwords. (Weir *et al.* 2009) developed a cracking algorithm using a Probabilistic Context-Free Grammars (PCFG) and a methodology for how to derive candidates for a password ordered by the highest probability.

They use the PCFG method only with the purpose of efficient password cracking. It is based on probabilities of occurrences of a given word with the lexical aspect of passwords. This strategy was considered as one of the best after the publishing, and even recent research is built on top of their method (Houshmand *et al.* 2015). However, this approach fails to understand relationships among words and their meaning.

In a publication focused on mobile security research, Jakobsson & Dhiman (2013) applied a different approach to assessing the password strength and the prediction. Their algorithm is able to take the password and decompose it using a parser and subsequently feed a model that predicts the probability of the occurrence.

In contrary to Weir *et al.* (2009), this approach permits to capture the structure of an alphabetical string. The disadvantage of this method is the inability to capture obvious patterns such as the insertion of numbers or appending symbols to the end of a string. A password "iloveprague123" would not be distinguished from "123iloveprague12345". This weakness is criticised by Veras *et al.* (2014). They developed NLP based methodology to account for this issue.

A few authors dedicated their time to study the semantic structure. Ur *et al.* (2013) decided to evaluate the effect of security policies on password quality and explain why it is happening. The team had a solid experience with research oriented on security and passwords, mainly having the relationship of policies and password strength as the centre of the research (Komanduri *et al.* 2011b; Kelley *et al.* 2012; Ur *et al.* 2012).

In this paper, they focused on words that compose the password and the relation among them. Their findings confirm that there are patterns beyond the well known ones, such as appending a number to a word from a dictionary. Given the results of the relation among chunks of text within a password, they believe the context-free analysis (Weir *et al.* 2009) discards potentially helpful information for modelling the patterns.

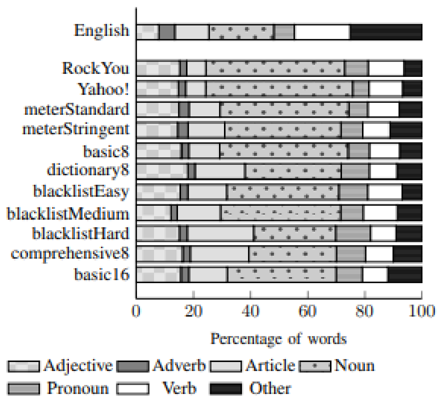
Additionally, the empirical evidence suggests users' alarming laziness when a system rejects their new password. Instead of creating a new strong password, they append a character to an existing one.

That suggests, multiple passwords of one user (evolution of passwords for one account) might be strongly similar, differentiating by a single character. The authors included several famous password breaches such as Yahoo! data leak or RockYou password database in the study. In total, they were working with a decent data set containing nearly 33 million passwords. They concluded that the possession of a part of the string increases the probability of guessing the rest of the password dramatically. A highly relevant outcome of the study is comparing the password corpus with the Corpus of Contemporary American English (COCA).

Authors compared the distribution of the sample data with the corpus and concluded that passwords were more likely than the English language to contain nouns and adjectives. On the other hand, significantly less likely to contain verbs or adverbs (see Figure 2.1).

That suggests that passwords are composed of short chunks of words rather than sentences. Speaking about distributions, they also compared how password chunks differ among themselves and in comparison with English. They claim a minor difference among password samples but a significant difference between English corpus and any password set. That suggests users tend to use similar language in passwords, no matter the website or service.

Figure 2.1: Distribution of pasts of speech for words in English in the samples and the English corpus.



Source: Ur *et al.* (2013).

The difference in the distribution of POS tags (i.e., the distribution of nouns, pronouns, verbs) was also investigated by Rao *et al.* (2013). The study's main goal was to examine the effect of grammatical structures on the vulnerability of long passwords. The study has three primary outcomes. First, they propose a framework to estimate the decrease in the search space[4] for guessing a password due to the presence of predictable grammatical structures. Second, they showed that the length of the string does not imply the strength of the password. Third, they proposed a technique for efficient password cracking using the estimated grammatical structure.

They used a brown corpus [5] to evaluate the search space given the sample data. They found that around 84% of passwords were generated similar to the brown corpus measured by the POS sequences.

That is in contrast with the findings by Ur *et al.* (2013). They claim, the structure of a password differs significantly from the English corpus. However, Rao *et al.* (2013) focused only on a subset of password breach containing solely long passwords. Thus, we might expect that the similarity of long passwords with natural language is higher. On the other hand, the findings based on the subset of data breaches could be hardly generalised to the whole population due to the selective sampling.

Their approach for password splitting and construction of a password space is based on POS sequences. Authors took the brown corpus that already contains POS tags for every password, preprocessed the text in terms of special characters and calculated all unique sequences of tags up to order n. Speaking in NLP terminology, for every sentence in the corpus, they generated n-grams of pars word and POS tag and then, they took unique values of n-grams up to order 10 to create a tag-rules object.

Considering a password "She runs fast", the corresponding POS tags would be "Pronoun Verb Adverb". A bi-gram of POS tags from this password would be (a) "Pronoun Verb" (b) "Verb Adverb", and there will be only one tri-gram "Pronoun Verb Adverb". They concluded that the ratio of observed unique tag-rules to all possible combinations is rapidly decreasing with the length of an n-gram.

For a bi-gram, the ratio is 99.92%. That means observed combinations of POS tags covers nearly all possible combinations. However, for a penta-

[4]The search space in the context of password guessing is defined as a set of all possible unique password values.

[5]Brown corpus is a well-balanced corpus of a language that contains a representative set of all grammatical structures for a given language.

gram, the ratio is only 46.5%. That means the structure of POS sequences can be used to decrease the search space of possible words for password guessing. Furthermore, some POS tags are more vulnerable than others (e.g., nouns are less likely to be used than pronouns). Their higher probability of occurrence gives that. As Veras *et al.* (2014) points, while semantic patterns are not discussed in the paper explicitly, it is evident that the semantic structure could reduce the search space of passwords even further.

This idea was further supported by Bonneau & Shutova (2012). They analysed patterns of human choice in a passphrase-based authentication system maintained by Amazon. They presented a corpus containing over 100 000 possible phrases to prevent users from using a simple and existing phrase as a password (e.g., "Extraordinarily Secure" is identified as easily guessable).

They enriched the data set with phrases from a natural language corpora of English. As expected, the conclusion was that the phrase selection is not random, and users tend to use well-known phrases composed by movie or book titles. This behaviour presents a threat to the security of user's accounts as they are easily guessable no matter their length.

Additionally, the evidence suggests users prefer simple noun bi-grams that are frequent in their natural language (English). However, they also conclude that the distribution of phrases in passwords is less skewed than in everyday language. That indicates that some users try to choose password randomly.

The analysis of POS n-grams reveals that in 13.3% of cases, a sequence of adjective and noun was found in the password. That is followed by a sequence of two consecutive nouns in 4.4% of the cases. Both numbers high enough to be used to decrease the search space for guessing.

This evidence of non-randomness of words further boosts the interest of studying the semantics of password as one would expect any well-defined pattern to decrease password security.

Chinese researchers recently published a study focused on comparing a lexical sentiment of passwords obtained from Chinese websites (Zeng *et al.* 2019). Up to our best knowledge, it is the most recent research that tried to examine the sentiment hidden in passwords using NLP.

The authors obtained three relatively large data sets for their study. First one, containing nearly 4.5 million passwords leaked from a Chinese Software Developer's Network. The second one, containing 4.8 million passwords that came from a social media website Renren[6] focused on sharing opinions, images

[6]http://reg.renren.com/

and messages. A large portion of the Renren users were students enrolled in universities or high schools. The last data set contained 9 million observations and came from a website with simple online games. Thus, they concluded that the combined data set should be representative enough.

Unfortunately, they do not provide if they had multiple observations per one user and how they dealt with it. Additionally, the number of observations was decent, but one might be concerned about the random sampling. One might expect that software developers and university students are highly educated in digital security and would choose more complex passwords than the rest of the population.

Furthermore, users of the gaming website (denoted as *T178*) will probably comply more with random sample assumptions. However, as it is a website dedicated to simple games, users might not care much about security. Thus, choosing simple passwords in contrast to relatively precious web sites such as banking, email, or social media such as Facebook (Hanamsagar *et al.* 2016).

Out of these 18 million observations, around 14 thousands of passwords are perceived as meaningful strings. This finding highly relevant to our study as it limits the set of passwords for the sentiment analysis. Surprisingly, in the first data set from the Developer's network, English words accounted for 25.9% among the sample. For the other two data sets, it was 17% and 15%, respectively. Those numbers are relatively high. We might expect to find English words in most developed countries where the command of the English language has become a standard.

In order to identify the sentiment in the passwords, authors had to segment the password into meaningful words. Fort his purpose, they used British National Corpus (BNC). They selected the top 15.000 words from this corpus. Additionally, they took 399 spellings from 4761 frequent Chinese words. They used 209 frequently used Chinese family names and a list of manually selected entities such as corporation names or internet slogans.

The splitting approach was the following: take a password and do a split if two neighbour characters differ in type (i.e., lower case, upper case, number of unique character). Next, they subset the language dictionary to match only the chunks identified in the first step. Afterwards, they identify the candidate with the highest coverage (i.e., cover the highest number of characters from the raw password). In case of a tie, they choose the candidate with a lower number of chunks (i.e., the least number of chunks). Finally, in case of a tie, they drop the observation.

The sentiment was studied from two perspectives—first, the polarity. Second, six kinds of emotions in the Ekman model. In general, they identified a sentiment word in 3.9% of cases in CSDN, 0.6% cases in RENREN and 0.4% in T178. That indicates heterogeneity among samples and possible different people's behaviour.

Positive words appeared more frequently than negative words. In the CSDN data set, researchers identified 10% of words as positive and only 4% as negative. Speaking about the Ekman emotions, joy was found to be the most frequent one, found in nearly 13% of the cases. The rest of the emotions was found in less than 1%. For the Chinese language, authors derive similar results from the most frequent and joy type of the Ekman list as the most popular one.

In conclusion, the authors demonstrated that positive words were more frequently used in passwords than negative ones. The joy emotion usage was the greatest among the Ekman list, and for the Chinese spelling, joy sentiment also dominated a considerable portion.

The polarity of language was a long time ago studied without any connection to passwords. The positive or negative emotion of a text was studied by Garcia *et al.* (2012). Their observations suggest that the frequency of words humans use is determined by the word length and the average information content. Furthermore, positive words tend to be used more frequently (in line with Pollyanna hypothesis (Boucher & Osgood 1969)). On the other hand, negative words tend to bear more information.

Zeng *et al.* (2019) are not the only ones who tried to understand the semantics of the passwords. A few years ago, Veras *et al.* (2014) published a study aiming to develop a framework for segmentation and semantic classification of passwords from a RockYou data leak. They argued that even after half a century of a password used in online and offline security, we still do not have a comprehensive understanding of the structure of the passwords and the rules users follow. A deep understanding is vital for a precise assessment of a strength of a password.

Their work implies three conclusions. First, they successfully demonstrated how NLP algorithms could be used to segment and classify passwords. Second, using the RockYou data, they estimated the most common semantic patterns. Third, they developed a modified Probabilistic Context-Free Grammar that captures the semantics, and they demonstrated how this information could be used for more efficient password cracking.

In that time, their method was estimated to guess 67% more Linked In passwords in the first 3 billion guesses than a state of the art algorithm. The authors aimed to capture the semantics to demonstrate how this knowledge could be misused for illegal purposes.

For the analysis, Veras *et al.* (2014) used the famous RockYou database of passwords containing over 32 million user accounts leaked in December 2009. The company (RockYou) was specialised in developing widgets for MySpace and implemented several applications for social networks such as Facebook. Thus, the sample of users consists mainly of users of social networks that registered in their platform and used their widgets or other services.

The passwords came in a plain-text version. The very first step was segmentation. That means splitting the string (password) into the most probable meaningful parts (segments). The first method was proposed by Jakobsson & Dhiman (2013). Their algorithm takes a combination of specialised and general dictionaries of a given language and uses the coverage, indicating how many characters were covered by the words.

Veras *et al.* (2014) follows this methodology. Furthermore, they suggest a strategy for identifying optimal split when more than one segmented version of a password is suggested. For example, having a password *catslightly*, we could segment it in cats-lightly or cat-slightly, both having the same coverage of 100%. To distinguish these two options, Veras *et al.* (2014) introduce higher-order N-gram[7] frequencies to disambiguate segmentation with equal coverage.

Both the coverage based segmentation and N-gram probability approach require a corpus to be estimated on. Authors employ two source corpora. The first one is a source corpus based on a collection of raw words used for the coverage based segmentation. The second one is a reference corpus, which is a collection of part-of-speech tagged N-grams. This reference corpus was used to select the most probable segmentation if the coverage based segmentation suggests more than one splits.

For the n-gram segmentation, they took a few steps to remove the noise and increase parsing speed. First, they removed all three-character words from the training corpora with a frequency less than 100. Second, they selected the top 37 two-character words based on frequency. Last, one-character words were eliminated to include only i or a.

Authors argue that the goal was to reduce the number of short and possibly

[7] In computational linguistics, an N-gram is a sequence of n items from a given sample of text or speech

rare words to increase the parsing speed without harming the accuracy. However, the authors do not provide an exact methodology on how these thresholds were chosen. They stated that the trimming was a result of observation of the dataset. Sadly, they do not even provide how the empirical selection was made.

The studied semantics cannot be covered by the COCA itself (i.e., the training corpora). Veras *et al.* (2014) also aimed to study Named Entities such as names, cities or surnames. Thus, they collect these lists by themselves from various sources. For example, names were derived from a US Social Security Administration dataset and cities were derived from the Geonames[8] with at least 15.000 inhabitants.

In addition to the segmentation mechanism, authors also discuss the problem of *mungling patterns*. Due to the lack of context, it is difficult to determine the correct segmentation of a password with 100% certainty. First, the corpora used for password parsing deviates from the standard language (as shown earlier). Second, users tend to make their password more secure by implementing simple patterns known as the *mangling patterns*. These patterns include replacement of characters, their deletion, concatenation and insertion (Jakobsson & Dhiman 2013).

As mentioned, the segmentation of password into words might yield several candidates as shown in Table 2.1. To identify the most probable candidate, the authors used the previously discussed N-gram probabilities based on the COCA corpus. They build an n-gram model up to order 3. The optimal password segmentation was chosen based on the probability of the segmentation n-gram. That seems to help segment short passwords composed of up to three words.

For long passwords composed of several words (i.e., more than three), this approach might fail to suggest the optimal password segmentation. Authors themselves admit that the decision of the n-gram order was made under the trade-off between accuracy and coverage. Higher-order n-grams might not be found in the corpus due to the sparse distribution, while using the recursive approach on sub-n-grams decreases the accuracy.

Authors applied the part-of-speech tagging procedure on the segmented passwords. They categorised the previously segmented words into categories (name, city, month) or a syntactic category (e.g., noun or pronoun). They created a custom tagger using the NLTK library[9] on the source corpus with

[8]https://www.geonames.org/

[9]NLTK is Natural Language Toolkit library for Python (https://www.nltk.org/index.html).

Table 2.1: Candidate segmentation for a sample password

Password		Segments				Coverage
Ageanyonebarks98	(A)	Anyone	barks	98		0.84
	(B)	Any	one	barks	98	0.84
	(C)	Anyone	bar	ks98		0.69
	(D)	Any	one	bar	ks98	0.69

Source: Veras *et al.* (2014)

an extension of the manually selected categories - name, city, surname, month and country.

The prepared data were fed into semantic categories. For this purpose, the authors used WordNet-based classification. For example, the words *car, auto, automobile, vehicle* would receive the same category: IS-A vehicle.

On top of this segmentation, authors developed a PCFG model on mangling rules and semantic patterns. It was shown that the model outperforms a benchmark set by Weir *et al.* (2009) dramatically. However, the guessing power is not of interest for this thesis. The password segmentation is of high relevance in this case.

This paper is a practical example of how to potentially split password into words using an extensive dictionary and statistical language models. Furthermore, it offers how the semantic patterns might be studied.

2.7 Hypothesised models

It was decided to focus in this thesis on two examples of poor password management. First, try to explain the similarity between a username and a password and second, attempt to find drivers of password reuse. Both practices might present a significant vulnerability to the security of an account.

Based on the literature review, it is expected that these two practices might be explained by a set of macroeconomic and microeconomic variables. Studies suggest that gender might play a significant role in determining the Password-Username similarity and the Password-Password similarity. Furthermore, it was found evidence that there might be cultural differences in terms of a language or a country.

It was also mentioned that the structural properties of a password might affect password management. It is believed that the character composition and the length might affect the two practices.

In addition to that, it is expected that one might be able to explain these poor password management practices by the environment the user is living in. One might argue, that the education influence the user's attitude towards password. Educated users might overall use better passwords. Similarly, users living in a country with a high level of digitisation might have higher security awareness and, thus, better passwords.

One more significant predictor might be the overall cybersecurity level in the country of the user. It might be expected that a high cybersecurity level also affects the user. They are educated about the potential risks of losing personal data. These users might prefer to protect their data carefully.

As the paper on sentiment in password suggests, the appearance of polarity in English or Chinese based passwords are not negligible. It might be interesting whether it is possible to identify the sentiment in passwords in other languages. Furthermore, if the presence of a sentiment affects the Password-Username similarity, it might be treated as an additional vulnerability to the account's safety.

Thus, this thesis aims to explain the following:

1. Password-Username similarity might be explained by Cybersecurity index, Democracy level, Mobile phone usage, Internet coverage, Literacy rates, password length, diversity of a password, gender and the sentiment.

2. Password-Password similarity, in other words, the reuse of passwords, might be explained by the Cybersecurity index, Democracy level, Mobile phone usage, Internet coverage, Literacy rates and gender.

Chapter 3

Methodology

This chapter is organised as follows. First, it is shortly described what software was used. Second, a thorough description of the obtained data. Third, what variables were used and how they were derived. Fourth, a detailed description of the sentiment assessment. Last, descriptive statistics giving an overview of the processed data.

3.1 Software and technology

The raw data for this thesis had more than 40 GB in text format. Initially, it was tried to use R with R studio, but severe problems with performance occurred. Even advanced libraries (e.g.: *data.table* package [1] or *stringr* package [2]) capable of handling large data were struggling with loading the text files and performing simple processing steps. R is operating in RAM and have difficulties processing files bigger than the memory.

Python was investigated as a second option. While it is also an in-memory based software, the performance improved. It was also tried to read the file line by line and thus, avoid the RAM size limitation. This method is very slow, even on SSD drive.

Apache Spark [3] was investigated as the next option and identified as a viable solution. The implementation in Python (PySpark [4]) was used, and the speed was exceptional. The technology is based on distributed computing, and while it is designed to run on servers managing large data, it performs decently

[1]https://www.rdocumentation.org/packages/data.table/versions/1.13.2
[2]https://www.rdocumentation.org/packages/stringr/versions/1.4.0
[3]https://spark.apache.org/
[4]https://spark.apache.org/docs/latest/api/python/pyspark.html

well, even on a conventional laptop. One of the advantages is the capability of efficient work with files larger than the RAM. Even though loading large files might take time, algorithms do not crash and finish successfully.

3.2 Data

Data in this thesis consist of a set of data leaks coming from different years and sources. In February 2018, a link on a large data set containing over 1 billion appeared on Reddit online forum. While the source is unknown, examining the individual files reveals several well-known data leaks.

Data were organised in small text files in a sizeable alphabetical tree structure. Altogether, data were divided into 2009 files occupying over 40 GBs on the drive. More than 1 billion observations were found in the sample.

A list of presented leaks was attached to the file. That included most of the well known data leaks such as LinkedIn leak (Kontaxis *et al.* 2013), (Veras *et al.* 2014), (Kamp 2012), RockYou leak (Yan & Chen 2018), (Fang *et al.* 2019), (Xu *et al.* 2017) or Yahoo! leak (Zhang *et al.* 2019), (Blocki *et al.* 2018).

Data came in the form of "username@provider.domain(s):password". The very first step in working with the data was to store them on an encrypted driver. Furthermore, data were anonymised, ensuring maximum security.

3.2.1 Initial data identification

The fundamental analysis was designed to be based on a country domain level. Thus, the raw text files in the tree structure were transformed into several parquets corresponding to the top-level domains (e.g., .com, .cz, .edu). This approach made it feasible to access and analyse data for a specific country fastly under the PySpark framework, discussed earlier.

At this point, while retrieving country data was straightforward, the observations were still in the raw, untouched format. It was necessary to separate the username, the provider and the password correctly. In most of the cases, there was a clear separator (i.e., a colon or a semicolon). Unfortunately, it was not always the case. Sometimes, a period was separating the email from the password, and in a minority of the cases, no specific symbol was separating those two structures.

A Regular Expression script was developed for the separation of the entities:

The logic of the script is following. First, look for any character other than

Figure 3.1: Regex used for the separation of the email and password

```
([\S^@]+)[@]([\w|\-]{1,})?[.|,](([\w|\-]+)[.|,])?\
(([\w|\-]+)[.|,])?(\w{2,10}|\-)[.|,]?[:|;|,|        | ]{0,1}(.*)
```

@ until @ is reached and consider it a username. Second, look for a text string with a period in the end and label it as the provider. Optionally, the next two chunks ending with a period are reserved for a longer provider or domain (e.g. .co.UK). After that, there should be a Top Level Domain (TLD) such as .com or .cz. Last, everything behind a colon, semicolon or comma is considered a password (including an empty set).

Table 3.1 reveals how the scripts performs the separation. Based on an empirical examination of the passwords, the incorrect split would be extremely rare, occurring in far less than 1% of the cases.

Table 3.1: A demonstration of the regex code

Raw observation	**Username**	**Provider**	**Domain 3**	**Domain 2**	**Domain 1**	**Password**
voj.tech@seznam.com:dogsandcats	voj.tech	seznam			com	dogsandcats
vojtech@fsv.cun.cz:!@#$%^&*()	vojtech	fsv	cuni		com	!@#$%^&*()
vojtech@example.co.uk:dogs	vojtech	example	co		uk	dogs
vojtech@fsv.cuni.co.uk:dogs	vojtech	fsv	cuni	co	uk	dogs

Another concern was the quality of the data, as some observations came in a suspicious format. For example, the observations were missing username, provider or domain and occasionally, password. The assumption was that a password had been compulsory, and thus, these observations with a missing password were treated as incorrectly decoded observations and were not included in the following analysis.

3.2.2 Data cleaning

The transformation of the raw string into its elements (i.e. a username, a provider, a domain and a password) produced several domains. Furthermore, the algorithm produced a large number of different providers that were not essential for this thesis.

At that time, there were more than 200 countries in the World. Due to the wide accessibility of the internet, it was reasonable to expect that each country had at least one Top Level Domain (TLD). Thus, one would expect approximately 200 TLDs in the data. Nevertheless, the data contained more than 200 000 different TLDs. This number was higher than expected because of several reasons.

First, a portion of observations contained typos and other forms of impurities. For example, if the user wrote *.comm* instead of *.com*, it would produced additional category. Frequency counts of the TLDs revealed that the impurity was not dramatic as there were far less than 5% of unexpected domains.

Second, some countries use more than one TLD. For example, the United States of America use *.us*, *.edu* for educational institutions and *.mil* for military organisations. Thus, the total number of domains was further increased by these special domains.

Third, there were Top Level Domains that were not assigned to any country. For example, *.biz* is a domain used by corporations around the World. Unfortunately, these domains are hard to relate to countries or nations, so their usage was limited.

Table 3.2 indicates the number of observations, users, domains and providers found in the parsed data. The raw data had nearly 1.5 billion observations. Furthermore, it contained more than 1.1 billion unique users, with more than 200 thousand domains and 14 million providers.

207 TLDs were identified as candidates for further statistical analysis of the users. More than half of the omitted data are related to *.com*. Thus, the resulting number of valid observations is considered to be a significant success.

Table 3.2: The basic statistics on the raw data and the valid sample

Type	**# observations**	**# users**	**# domains**	**# providers**
raw	1 403 267 127	1 153 989 209	228 136	13 738 123
sample	491 617 687	394 894 148	207	5 043 234

Table 3.3 indicates number of observations per domain in the sample. Domains in red were excluded from the country analysis as it was not feasible to match them with a country. Nearly all modestly frequent domains were retained, and only non-country related domains or unknown domains were omitted. A complete table indicating the raw data counts can be found in the Appendix in Table A.1.

At this point, the sample contained only valid domains. However, further assessment of the quality of the data was necessary to perform. There might have been suspiciously long passwords or usernames, and invalid values might have populated the fields.

Table 3.3: Observations per domain with omitted domains in red

#	TLD	Count	#	TLD	Count	#	TLD	Count
1	com	844 200 121	11	edu	6 851 080	21	hu	2 411 229
2	ru	226 594 848	12	jp	6 186 806	22	tw	1 982 277
3	de	63 271 746	13	br	5 813 716	23	mx	1 950 652
4	net	50 048 314	14	es	5 702 419	24	id	1 862 310
5	fr	44 986 923	15	ca	4 806 575	25	at	1 565 892
6	uk	26 644 014	16	ua	4 361 824	26	sk	1 505 968
7	it	24 675 740	17	au	3 918 009	27	be	1 485 327
8	pl	13 261 771	18	org	3 596 060	28		1 473 927
9	cz	7 653 736	19	nl	3 342 572	29	za	1 343 648
10	cn	7 200 950	20	in	3 200 151		*Other*	*31 368 522*

The assessment of the quality was therefore done in two steps. First, it was verified that the value has expected format and second, it was evaluated whether the value itself is meaningful.

The password The password field was considered the most problematic one. Based on the sample, the maximum length of a password was more than 120 characters, a suspiciously large number. If an average English word had around five characters, that would, on average, imply approximately 24 words in one password. Because of that, an investigation of the length of a password was performed.

The data frame was sorted according to the length of a password. It was manually evaluated whether it contains meaningful data or not. First, it was attempted to identify patterns that math an incorrect value. Fortunately, a significant part of the incorrect observations was either type of a connection string (i.e., strings that describe a connection to an account through an address, username, password and a protocol) or hashed version of a password. Based on this empirical evaluation, Table 3.4 reveals what prefixes were considered non-password related.

In addition to these connection strings, it was attempted to identify observations that failed to be decrypted. These observations frequently appeared as a long random sequence of numbers and letters. Optionally, they appeared with a prefix. The prefix could vary depending on the technology, where the credentials had been stored. Table 3.4 indicates what patterns were considered as a hash format.

The password cleaning approach was based on empirical observation, and

Table 3.4: Passwords in hashed or invalid form

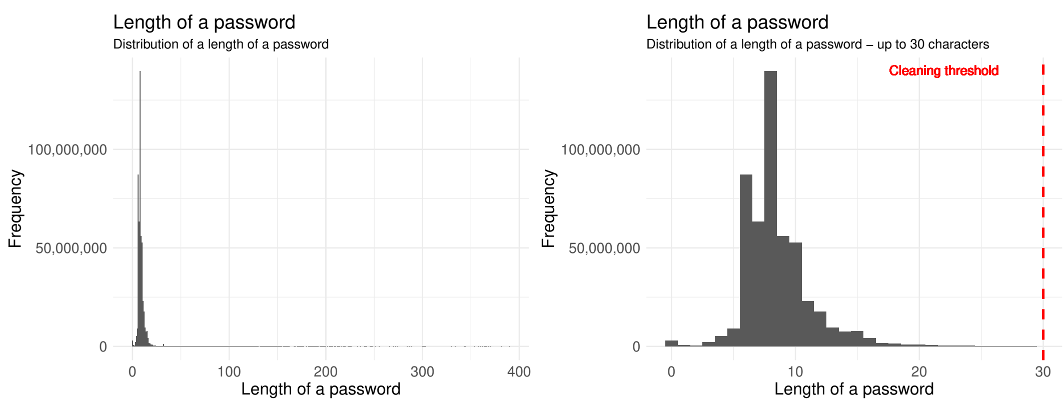
Suspicious pattern	Example of raw value
ssl://	example:||ssl://smtp.example.cz||465||
$HEX[	$HEX[11111111111111111111111111111111]
find_pass=	find_pass=111111111111111111111111111
:||imap.	example:||imap.example.de||25||example|
:||smtp.	example:||smtp.example.de||25||example

eventually, it might introduce some bias. Nevertheless, it was done with solid carefulness, as shown in Table 3.4. The approach could be further improved by a statistical analysis of the distributions of prefixes and lengths of the passwords. Furthermore, that could be enhanced by taking the TLDs into account as some of the domains (e.g., *.ru*) dominated the number of hashed passwords.

While this cleaning step dramatically reduced the number of broken records, the sample still contained a few suspiciously long passwords. Unfortunately, the application of filters based on the connection string structure and hashed structure of the password were not sufficient. Thus, it was decided to evaluate the remaining observations manually one more time and find a length of a password where incorrect passwords begin to overlap with the correct records.

Figure 3.2 reveals the distribution of a password length in the sample. On the left side, one can see that the distribution is highly skewed, and the chart is hugely shrunk due to a few enormous values. On the contrary, the chart on the right side shows the same distribution with an interval ranging from 0 to 30 characters.

Figure 3.2: Distribution of length of a password



Almost all passwords in the interval above 40 characters were sequences of numbers and letters. That suggests these passwords were another hashed forms of the original password strings. Therefore, all these rows were considered redundant for this thesis and thus, were eliminated. This decision is supported by the unexpected length and by the composition of the strings. This decision reduced the data by less than 1%.

The resulting data set was considered to be cleaned in terms of passwords and ready for further analysis. The discarding of suspicious values was done with deliberation, and it was not expected that it introduced significant bias to the data. It was believed the noise was reduced dramatically.

The username The username was used in the analysis to estimate the similarity between a password and a username. Thus, an assessment of the quality of the variable was performed as well. While it was not known what might be the maximum length of a username, we might expect the upper boundary would not be more than a couple of dozens of characters without loss on generality. Unfortunately, the maximum length of a username in the sample was almost 400 characters.

Figure 3.3: Distribution of length of a username

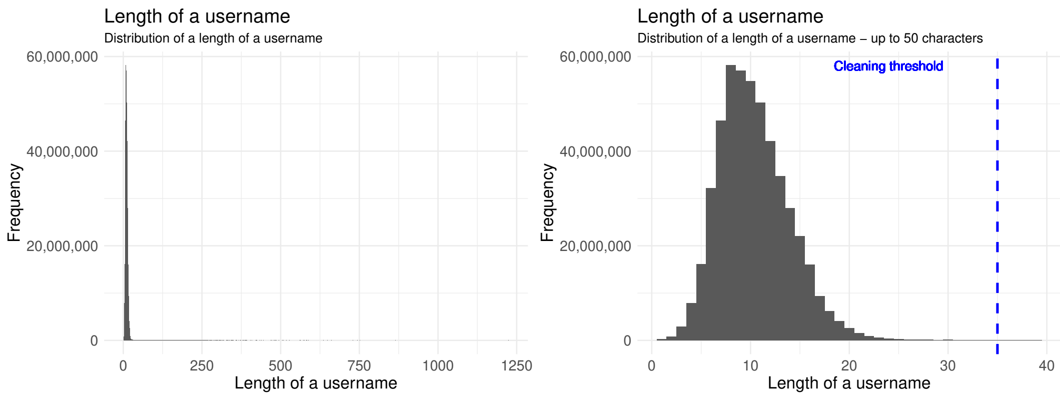


Figure 3.3 indicates the distribution of a username's length. The chart on the left reveals the whole distribution, indicating several suspiciously large values (e.g., 500). On the right, one can see the same distribution on the interval from 0 to 40. The mean is around ten characters, and the distribution is well bell-shaped.

The data ware sorted by the length of the username. Afterwards, rows were empirically evaluated, finding a threshold of length where the username start to be valid. Surprisingly, up to our best knowledge, all of the observations with at least 50 characters were random sequences of numbers or special characters.

Furthermore, below approximately 50 characters, observations started to be a meaningful text.

As the proportion of these meaningful strings was still decently low at length 50 (less than 1 out of 10), the cutting threshold was set to be even lower. Around 25 characters in length, more than half of the observations were identified as valid. Thus, to relax this threshold and include as much as possible correct records while discarding noise, the limit was set to be 35. That is, any username longer than 35 characters were eliminated.

On the contrary to the length of a password, the username's length seemed to have a higher mean and median. That suggests that while users have decently long usernames, their passwords are noticeably shorter.

The domain The main domain is the Top Level Domain (e.g., *.cz*). A TLD can have a subdomain, which is called a second-level domain. These second-level domains can correspond to several third and lower-level domains. The regular expression code captured this fact. The maximum supported order was the third one. Nevertheless, only the TLD was used in the analysis and thus, potential minor discrepancies in the derivation of lower-level domains would not impact the country assessment.

The provider Users might create an account at various companies (e.g., Google.com, Seznam.cz or yahoo.com), and these companies are called providers. A provider could be a piece of valuable information. It would be interesting to examine how users' behaviour differs across a banking and shopping account.

Unfortunately, the email address is frequently used as a username for a completely unrelated web service. Thus, we might know that a person has a bank account, but there is no guarantee that the sample's account information corresponds to the identical service. For example, a customer of Google might register at Amazon with a Gmail address, and it would not be feasible to identify the actual provider.

If it would be possible to identify the website where the email was used (regardless of the parsed provider), it might be possible to segment the data by the importance of websites. It might be expected that users use stronger passwords for Internet Banking while having weak passwords for an eBay account.

At this point, the data was considered to be cleaned and ready for further analysis. The suspicious length of a password and username was corrected, and the domain was cleaned.

Cleaned sample Table 3.5 reveals how many observations per TLD were found in the cleaned data. The TLD can be directly linked to a country. As one can see, the most frequent country was Russia, followed by Germany, France, UK, Italy and Poland. The least frequent countries are either small or developing countries such as Kongo, Chad and Guyana.

As the number of inhabitants varies among the countries, a column indicating the country's population is presented as well. Furthermore, the ratio indicates the number of observations per number of inhabitants in the country. For developed countries, the ratio is high, reaching more than 1.5 for Russia. That means there are more than 1.5 accounts per one inhabitant, on average. For Germany, the ratio is lower, however, still reaching a decent value of 0.7. On the contrary, for some countries, the ratio is unquestionably lower, falling to 0.0001. A complete table indicating the full sample ratio can be found in the Appendix (see Table A.2).

Table 3.5: The most and the least frequent countries in the sample

TLD	Country	Observations	Population	Ratio
ru	Russian Federation	224,529,891	144,478,050	1.5541
de	Germany	63,080,765	82,927,920	0.7607
fr	France	44,762,457	66,987,244	0.6682
uk	United Kingdom	26,471,527	66,488,990	0.3981
it	Italy	24,558,155	60,431,284	0.4064
pl	Poland	13,209,188	37,978,548	0.3478
cz	Czechia	7,601,246	10,625,695	0.7154
cn	China	7,167,071	1,392,730,000	0.0051
edu	usa -edu	6,767,896	327,167,420	0.0207
jp	Japan	6,170,003	126,529,100	0.0488
⋮	⋮	⋮	⋮	⋮
sl	Sierra Leone	671	7,650,154	0.0001
bj	Benin	577	11,485,048	0.0001
cg	Congo	541	5,244,363	0.0001
gy	Guyana	521	779,004	0.0007
km	Comoros	455	832,322	0.0005
gn	Guinea	291	12,414,318	0.0
iq	Iraq	173	38,433,600	0.0
cw	Curacao	146	159,849	0.0009
td	Chad	121	15,477,751	0.0
gw	Guinea-Bissau	67	1,874,309	0.0

The Ratio is calculated as the number of observations over the number of inhabitants.

This thesis also focused on the reuse of passwords by a single user. Table 3.6 reveals the number of distinct users per country and the number of users that appeared at least twice in the sample. The average share of recurrent users was around 10%.

That was is a significant portion in relative terms, but on the other hand, it was a decent number of observations in absolute numbers. There were dozens of millions of observations of recurrent users. The complete list is presented in the Appendix in Table A.3. In total, there were 491 617 687 observations related to 394 894 148 distinct users. Furthermore, out of this number, 55 047 171 users appeared at least twice.

Table 3.6: Distribution of account information per country

Country	**Observations**	**Providers**	**Users**	**Rec. users**
Russia	224,529,891	539,789	178,008,405	25,516,318
Germany	63,080,765	744,031	50,429,761	4,885,056
France	44,762,457	169,911	34,347,575	6,684,037
United Kingdom	26,471,527	1,025,452	22,014,131	3,095,395
Italy	24,558,155	331,975	18,632,106	3,751,850
Poland	13,209,188	130,334	10,462,959	1,672,662
Czechia	7,601,246	92,235	6,094,203	823,177
China	7,167,071	107,228	6,378,743	650,001
USA - edu	6,767,896	38,923	6,045,968	591,797
Japan	6,170,003	102,210	4,948,527	725,173
⋮	⋮	⋮	⋮	⋮
Sierra Leone	671	200	643	23
Benin	577	164	533	38
Congo	541	247	455	73
Guyana	521	216	478	27
Comoros	455	177	358	83
Guinea	291	72	253	31
Iraq	173	73	152	11
Curacao	146	80	132	14
Chad	121	69	110	10
Guinea-Bissau	67	50	61	6

Rec. users stands for *recurrent users*, users that appear at least twice.

3.3 Model construction

In this section, the author presents what variables were identified as necessary for the analysis and how they were derived and calculated. First, an overview of the hypothesised models is given and second, a detailed description of the variables is presented. The statistical properties of the variables are described in the next section.

3.3.1 An overview of the variables in the models

Following the literature review, two hypothesised models related to the human's password management attitude were stated. First, it was aimed to understand the irresponsible attitude to password management by explaining the similarity of a username and a password. Second, it was intended to understand the poor password management by comparing the similarity of passwords of one user.

It was believed that at least three following categories affect the decision-making of the users. Table 3.7 reveals a summary of the groups, descriptions, and examples of the variables. The first group was related to the written form of the password. That means studying the password length or range of characters used. The second group was describing the user itself; one's personal skills and properties. The last group was capturing macroeconomic properties, similarly affecting a group of users. The following subsection describes the identified factors.

3.3.2 Description of the variables

The password length Longer passwords are harder to remember. Simultaneously, the length of a password is generally associated with a higher quality of a password. It might be assumed that users that choose longer passwords are aware of the security and would be less prone to derive their passwords from their usernames.

Password quality is hard to measure, and as discussed in the literature review, there are multiple views on the measurement of the strength of a password. The length of a password is not the best measurement of the strength of a password but serves well as an indicator of the user's general quality and attitude. There is no doubt that a string containing two characters is significantly less secure than ten characters long password regardless of the source character set.

Table 3.7: Description of variable categories used in this analysis

Category	Description	Example
Password derived features	Features, that are derived from the raw version of a password and describe the properties of the string itself. Describes form of a password and omit lexical content and statistical properties.	Length of a passwords
Microeconomics variables	These variables are directly linked to the users and describe their attitudes and systematic difference in behavior. These variables should describe the user very accurately.	Education level
Macroeconomics variables	It is believed the security awareness differ among countries. These variables tries to capture fixed effects induced by country differences.	Cyber security level

The effort It would be interesting to measure the effort made by people related to password management. Even though a password is long, it might be concluded that the user did not make an effort in the management as all characters are lower case letters and together form a well-known phrase. It is believed that the effort made by people negatively affect the similarity between a username and a password and the reuse of a password as well. In order to capture this attitude, the effort variable was introduced.

Passwords can be composed of four general types of characters: lower case letters, upper case letters, numbers and special characters. In general, a wide selection of characters used impacts positively the password quality. On the other hand, a password composed of all four character groups is more challenging to remember than a password composed of lower case letters only.

To express this effort, it was decided that the effort would indicate how many groups were covered in the password. For example, a score of 3 means that password might be composed by a lower case, upper case and by a number. Table 3.8 demonstrates how this indicator works through a set of examples. It was expected that if users had the effort of 4, they cared about their accounts' security and should have low similarity between the username and a password.

Gender It was believed that there were several factors affecting password management describing the individuals. For example, education, security aware-

Table 3.8: Demonstration of the Effort indicator

Effort level	Example of a password
0	""
1	"hello", "1234", "!@#$", "CAT"
2	"cat1", "CAT1", "cat$", "caT"
3	"CAt1", "CAT1$", "cat1$"
4	"Cat1$"

ness and age. Educated people might be aware of the risk they undertake with poor passwords, and users with high-security awareness might be significantly more careful when managing passwords. Older people might struggle with memory and thus, prefer shorter passwords, while young users might choose long, high-quality passwords that are harder to memorise.

As discussed in the literature review section, the effect of gender on password management was not apparent, and existing reports made conclusions based on relatively small samples. Thus, it was expected to contribute to the discussion by bringing the analysis results on extensive data.

The gender had to be derived from the observations as there was no information about the users. The strategy was to obtain an extensive list of names in multiple languages and try to match it with usernames in the sample. www.behindthename.com containing names in dozens of languages was the primary source for this analysis. A script for web-scrapping was written in python, and all relevant names were downloaded according to the domain list from the sample.

All together, it was managed to scrape 36 332 names corresponding to 149 countries. However, there were only 20 949 unique names. That suggested that a significant portion of names was shared among multiple languages.

Table 3.9 reveals the number of names per language that was extracted. A few names were used for both women and men. For example, in the Indian language, 97 names are used by males and females, while 446 names are used by males solely and 295 by females. Fortunately, for most languages, the share of names used by both genders is relatively small.

In the analysis, the user's language was unknown and had to be derived from the data. First, the country was determined through the domain and second, the language was assessed based on official languages in a given country. A country might have multiple official languages or unofficial language spoken by a non-negligible share of the population, and the gender identification method

Table 3.9: Number of names per language

Language	F	F/M	M	Language	F	F/M	M
african	197	174	200	kazakh	27	0	32
akan	14	8	12	khmer	9	11	3
albanian	32	1	50	korean	40	52	49
amharic	13	4	12	kurdish	10	5	19
ancient	431	12	1319	latvian	121	0	99
arabic	350	53	527	lithuanian	115	0	116
armenian	39	4	72	macedonian	151	3	174
azerbaijani	39	1	46	malayalam	41	7	84
basque	88	3	91	mongolian	19	2	9
belarusian	45	0	50	ndebele	12	6	10
bengali	36	10	132	nepali	18	11	57
berber	5	0	4	norwegian	355	10	346
breton	17	4	37	occitan	10	1	7
bulgarian	178	6	184	odia	0	0	15
catalan	86	1	97	pakistani	46	17	132
croatian	299	6	288	pashto	1	1	17
czech	248	5	202	persian	111	16	144
danish	315	10	301	polish	265	2	273
dutch	402	30	435	portuguese	309	8	374
egyptian	3	0	19	punjabi	6	3	43
english	2276	310	1466	roman	104	4	196
esperanto	32	0	21	romanian	145	4	155
estonian	55	0	52	russian	230	14	304
filipino	10	0	7	sardinian	7	1	11
finnish	289	4	282	scandinavian	80	0	150
french	506	38	425	scottish	104	12	224
galician	17	1	33	serbian	200	4	208
ganda	4	1	5	slovak	174	0	136
georgian	74	0	118	slovene	213	6	221
german	511	19	470	spanish	667	22	675
germanic	204	3	631	swahili	15	1	12
greek	110	2	359	swedish	357	8	322
hawaiian	30	24	22	tajik	2	0	14
hebrew	202	67	193	thai	16	4	12
hindi	197	40	252	tibetan	0	17	2
hinduism	56	22	88	tswana	5	10	3
hungarian	264	2	216	turkish	276	33	339
chinese	13	86	7	turkmen	6	0	8
icelandic	108	2	103	ukrainian	108	4	95
indian	295	97	446	uzbek	14	0	24
indonesian	33	25	57	vietnamese	29	19	28
irish	234	19	363	welsh	130	14	187
italian	504	11	552				
japanese	163	44	175				

should account for that. Table A.9 in the appendix indicates what languages were identified per domain.

The exact procedure for the name derivation was following. Data was evaluated row by row. For every username, it was attempted to find all names of the language identified by the domain in the username. If a name was found, the gender was derived using the scrapped table. If multiple names were found, it was checked whether all names are of the same gender. If there were the same, gender was estimated. Table 3.10 presents an example of how this procedure worked.

Table 3.10: Demonstration of the gender identification

Username	**Identified Names**	**Gender**
abcdefg	None	Unknown
jessica123	Jessica	Female
oliver123harry	Oliver, Harry	Male
oliver_and_jessica	Jessica, Oliver	Female/Male

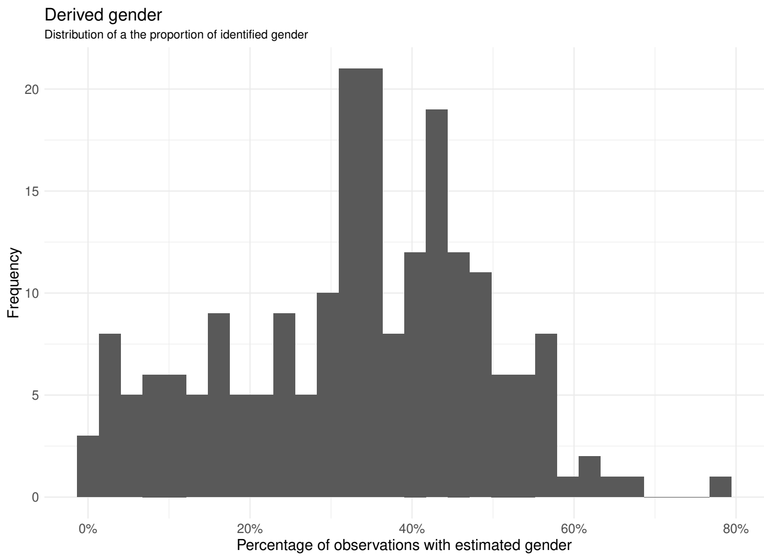
In total, 75 910 382 male observations and 75 910 382 female observations were identified. That means it was managed to derive the gender for nearly 30% of the observations. Furthermore, in 4% of the observations, a name was found, but the gender was ambiguous. Either due to multiple occurrences of names (i.e., both female and male) or because the name was used by both genders.

Table A.4 presents an overview of identified female and male names per country. While for some countries, the number of identified gender of users was decently high (e.g., Finland, Philippines, Belgium, Ireland), for some countries, the number was significantly lower (e.g., Greece, Somalia, Uzbekistan). Figure 3.4 reveals the distribution of the percentage of successfully identified names from the data.

As the given name extraction was not successful for all the observations, it was attempted to develop another extraction method. Users might put their surname in the username instead of the given name. However, these observations were hard to decode. In some languages, such as the Czech, female surnames have a specific suffix and, thus, are feasible to decode. Unfortunately, this method does not apply to all countries.

It was also considered to search for a name in the password. Empirically, a decent number of passwords contain a name. However, it is impossible to

Figure 3.4: Identification of gender



guarantee that the name found in the password corresponds to the user. It might be the name of a friend, a spouse or a famous person. Thus, it was not taken into consideration for this thesis.

Education As discussed, it was believed that education affects password management attitude. People with a good education might be well aware of the risks associated with poor password management, while uneducated people might be irresponsible regarding password creation and storage.

Ideally, one would know the education level of every single user in the sample. Unfortunately, this information was not provided, and the estimation of user-level was hardly possible. Nevertheless, the interest was also to estimate country-specific effects, and because of that, a country level approximation of the education might indicate the overall importance of education on password management and creation well.

Literacy rate of adults was identified as a suitable approximation of the education level in a country. The vast majority of the data was taken from The UNESCO Institute for Statistics [5]. This data source covered more than 90% of the required countries. For the missing countries, Macrotrends [6] were used as a second source.

[5] http://uis.unesco.org/
[6] https://www.macrotrends.net/

Freedom As discussed, it was expected that the level of freedom and the perception of the security might be related to password management and creation. To approximate this, two variables following the literature review were identified—first, a democracy level and second, the cybersecurity level.

People in non-democratic countries might feel followed by the regime, and thus, they might be more careful with their electronic activities. For example, China creates a social index measuring how loyal one is to the party (Zeng 2016). Private communication between individuals in disagreement with the party might negatively affect their life. Thus, it was expected that they look for secure manners of communication which also includes better passwords. Thus, the level of democracy was chosen as one of the factors.

The democracy index provided by The Economist Intelligence Unit from 2016 was chosen [7], and it covers all countries in the sample.

In addition to the democracy level, the Cybersecurity index was included. People living in countries with a high level of cybersecurity might be aware of the potential data breach and might chose better passwords.

Thus, the Global Cybersecurity Index issued by the International Telecommunication Union (Cravo *et al.* 2019) was chosen. The authors created 50 questions and collected answers from the countries. Furthermore, they enhanced the information by publicly available data and created an index to estimate countries' cybersecurity level. This index covers all countries in the sample.

Digitisation It was believed that the general awareness of cybersecurity might be affected by the overall level of digitisation (i.e., how are people used to use electronic devices). It was expected that nations used to use various electronic devices and services might be more aware of the danger implied by poor security attitude. Either they might have experienced a loss of credentials, heard about it from other users or read about it from the newspaper.

On the other hand, people in countries with low digitisation might not be aware of the implications of the reluctant attitude towards security and might use poor passwords.

The digitisation is not easy to measure. As a proxy, two variables were identified. First, the level of internet coverage and second, the number of users with access to mobile phones.

The internet usage was downloaded from the World Bank, and the source was the International Telecommunication Union. The variable indicated the

[7] http://felipesahagun.es/wp-content/uploads/2017/01/Democracy-Index-2016.pdf

percentage of individuals using the internet. According to the official definition, internet users are individuals who have used the internet (from any location) in the last three months, while the internet can be used via a computer, mobile phone, personal digital assistant, games machine or digital TV. The usage was available for all countries in the sample.

The mobile usage was the mobile cellular subscription per 100 people. In simple words, it is approximately the number of mobile phones per 100 people. Nevertheless, the official definition is following:

> Mobile cellular telephone subscriptions are subscriptions to a public mobile telephone service that provide access to the PSTN using cellular technology. The indicator includes (and is split into) the number of postpaid subscriptions and the number of active prepaid accounts (i.e. that have been used during the last three months). The indicator applies to all mobile cellular subscriptions that offer voice communications. It excludes subscriptions via data cards or USB modems, subscriptions to public mobile data services, private trunked mobile radio, telepoint, radio paging and telemetry services.

The source is the International Telecommunication Union, and data were available for all countries in the sample.

Password similarity The first hypothesised model captures why people make their accounts vulnerable by using similar passwords and usernames. It was necessary to build an approach for measuring the similarity.

First, it was tried to use the longest common sequence (LCS) measure. LCS could be used to calculate the similarity between the username and the password. However, this measure does not take into consideration the length of a password. Consider a username *vojtech123* and a password *vojtech999*. The longest common sequence is *vojtech*, and its length is 7. The length of the LCS and the username (or password) length are on a similar scale. However, for a username *vojtech123* with a password *vojtech123456789*, the longest common sequence become hard to accept as it does not take into consideration the length of the string.

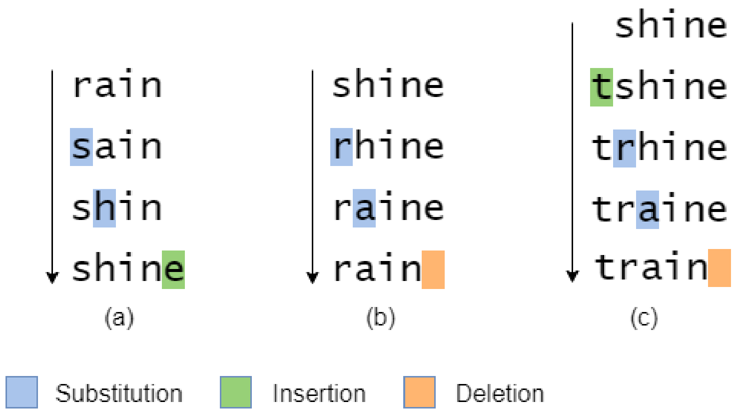
Next, advanced measures for string similarity were investigated, and two potential candidates identified. The Levenshtein distance and the Hamming distance.

The Hamming distance indicates the number of character positions (index) in which the two characters from the two strings are different. This approach, however, is suitable for strings with equal length. Given the nature of the data used in this thesis, that is something hard to expect. In most cases, the length of the username and password differs, so this distance measure is not suitable.

On the contrary, the Levenshtein distance supports the comparison of strings with a different length. Consider three operations: insertion, deletion and substitution. These operations can be used to transform one string into another. There are several ways how to achieve this transformation. The Levenshtein distance is defined as the minimum number of operations required to make the two inputs equal[8] Lower the number, the more similar are the two inputs that are being compared.

Consider two words, *rain* and *shine*. In order to transform *rain* into *shine*, one has to substitute r by s at the beginning, a for h at the second position and append e at the end of the string. This transformation results of the Levenhstein distance of 3. The transformation is also demonstrated on Figure 3.5

Figure 3.5: Example of the Levenhstein distance measurement



Source: https://devopedia.org/levenshtein-distance

Table 3.11 compares the measures for the three described approaches.

Table 3.11: Comparison of different distance measures

Username	Password	LCS	Hamming distance	Levenhstein distance
vojtech	vojtech	7	0	0
vojtech	vojtech123	7	NA	3
vojtechne	vojtech12	7	2	2
vojtech	hamster	0	7	7

LCS stands for *Longest Common Sequence*.

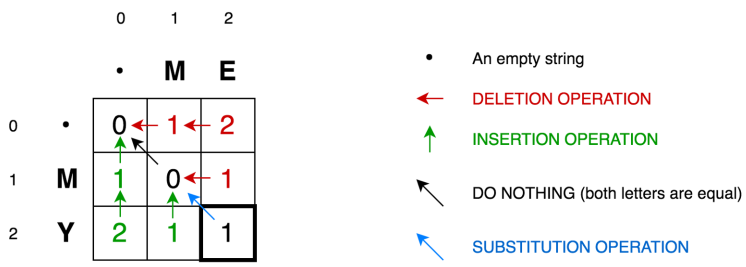
[8] https://devopedia.org/levenshtein-distance

Equation 3.1 is a formal notation of the Levenhstein Distance in a recursive form. a and b indicates two strings, i and j are indexes of characters in the strings a and b. The algorithm can be demonstrated in a matrix form, having the first word on x axis and the second word on the y axis. Figure 3.6 demonstrates the possible transformations (i.e. deletion, insertion, substitution) in the matrix form.

$$\text{lev}_{a,b} = \begin{cases} max(i,j) & \text{if } min(i,j) = 0, \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases} \tag{3.1}$$

The position (2,2) in the matrix indicates the number of moves that have to be done to perform the transformation. Having the pair *Me* and *My*, one only needs to substitute the letter e with y. Thus, the path has a length of 1. Nevertheless, it is not difficult to observe that there are multiple paths connecting positions (0,0) and (2,2). The goal is to find the shortest path, and the Levenshtein Distance algorithm finds this number. Figure 3.7 demonstrates the Levenshtein Distance and the moves on a larger example.

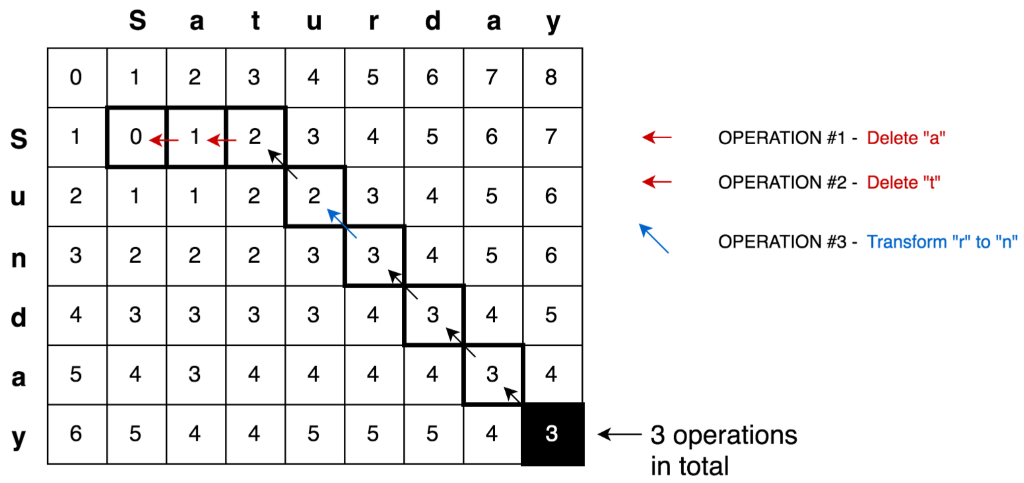
Figure 3.6: Moves of the Levenhstein distance in matrix form



Source: https://itnext.io/dynamic-programming-vs-divide-and-conquer-2fea680becbe

In conclusion, the Levenshtein Distance was used to calculate the similarity between a username and a password and among two passwords as well.

Figure 3.7: Moves of the Levenhstein distance in matrix form (Example 2



Source: https://itnext.io/dynamic-programming-vs-divide-and-conquer-2fea680becbe

3.3.3 Model framework

In conclusion, two base models were built using the described factors. The first model explains the similarity between a username and a password, and the second model, explaining a password's derivation from a previous one. Table 3.12 indicates the abbreviations of the variables for further reference.

Linear Regression In both model families, the explained variable was a whole number ranging from 0 to 30. That many levels of the dependent variable might be well fitted by the standard Linear Regression. That model choice would allow for easy fitting and interpretation of the results.

Unfortunately, two essential assumptions would not be met. First, the dependent variable is not continuous and second, the change in the target from 1 to 2 is not the same as from 29 to 30. The change from 1 to 2 has more significant security implications than the latter one.

Multinomial Logistic Regression As the target is described as the number of required modifications of a password (or a username), it can hardly be considered a real number. There is nothing such as 2.5 modifications. Thus, a family of Generalised Linear Models might be considered. The 30 distinct values could be treated as a multinomial output. That could overcome the issue of non-continuous variable and different distance between numbers.

Unfortunately, Multinomial Logit would not retain the ordering of the predicted variable. Clearly, the order of the number of modifications is essential.

Ordered Logit One of the models that allow for retaining the order of discrete dependent variable is Ordered Logistic Regression (McCullagh 1980) known as Proportional odds model or Parallel lines model. This model allows working with a set of non-numeric outcomes that follow a specified order. An example would be a customer's satisfaction: low, medium and high, where the order of the choices is clear.

Having k ordered target categories and m independent variables, the Ordered Logit can be denoted as

$$log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = log\ \frac{P(Y \leq j)}{1 - P(Y \leq j)} \tag{3.2}$$

for $j \in (1, ..., k-1)$ and defined as

$$log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \alpha_j + \beta_1 X_1 + \cdots + \beta_m X_m \tag{3.3}$$

for $j \in (1, ..., k-1)$. That can be expanded to the individual equations as

$$\begin{aligned}
\log\left(\frac{P(Y \leq 1)}{1 - P(Y \leq 1)}\right) &= \alpha_1 + \beta_1 X_1 + \cdots + \beta_m X_m \\
\log\left(\frac{P(Y \leq 2)}{1 - P(Y \leq 2)}\right) &= \alpha_2 + \beta_1 X_1 + \cdots + \beta_m X_m \\
&\vdots \\
\log\left(\frac{P(Y \leq k-1)}{1 - P(Y \leq k-1)}\right) &= \alpha_{k-1} + \beta_1 X_1 + \cdots + \beta_m X_m
\end{aligned} \tag{3.4}$$

where α_j is the intercept related to the cutoff j and β_i is a coefficient corresponding to the i-th predictor, which is identical across the equations. Individual equations represent the probability of being at or below the textitj-th category. Furthermore, a reference to the cutoff j indicates where the dichotomisation of the categories into the binary values 1/0 is performed. In other words, categories lower or equal than j are treated as one and the rest of the categories as 0.

One of the main assumptions of the model is the proportional odds assumption. As Williams (2016) points out, if the assumption is met then the odds ratios will remain the same regardless of which of the collapsed logistic regression is estimated. From a different point of view, that also implies that the β coefficients are invariant in terms of the cutoff *j*. That means, for example, that for a variable X_1 there is only one estimated coefficient β_1 for all *k-1* equations.

In order to make the interpretation of the model more intuitive, the cutoff j is sometimes considered in the reversed order. That means, that instead of modelling $P(Y \leq j)$, researchers model $P(Y \geq j)$. That allows for interpreting the β coefficients in the same direction as the cutoff *j* increases. In other words, an increase in a variable with a positive β estimate increases the likelihood of falling into a higher category, indicating a positive effect on the outcome.

Formally, one can interpret the effect of a coefficient β_i as the increase in log odds of at or above a category associated with a one-unit increase in X_i, holding the rest of the explanatory variables fixed.

Long & Freese (2014) offer an intuitive interpretation of the Ordered Logit model derived from the proportional odds assumption. Researchers are frequently used to regress a continuous variable Y on some variables *X*. For a moment, one might consider Y not to be continuous but rather a collapsed variable of an unobserved latent variable Y^*.

For example, let us assume the case of a questionnaire where people can choose one of the ratings low, medium or high. As the X variables changes, the latent variable Y^* changes as well, and respondents eventually cross the thresholds on the latent variable Y^*. That means they move to a different rating as well.

Despite the popularity of this model, empirical observation suggests that the proportional odds assumption is frequently violated (Long & Freese 2014). If the assumption is violated, then the generalised ordered logistic regression might be considered instead.

Generalised Ordered Logit The Generalised Ordered Logit is similar to the Ordered Logit. The main difference is that it allows the β coefficients to differ across the cutoffs. Thus, the model can be described by the set following equations

$$
\begin{aligned}
log\left(\frac{P(Y \leq 1)}{1 - P(Y \leq 1)}\right) &= \alpha_1 + \beta_1^1 X_1 + \cdots + \beta_m^1 X_m \\
\log\left(\frac{P(Y \leq 2)}{1 - P(Y \leq 2)}\right) &= \alpha_2 + \beta_1^2 X_1 + \cdots + \beta_m^2 X_m \\
&\vdots \\
\log\left(\frac{P(Y \leq k-1)}{1 - P(Y \leq k-1)}\right) &= \alpha_{k-1} + \beta_1^{k-1} X_1 + \cdots + \beta_m^{k-1} X_m
\end{aligned}
\tag{3.5}
$$

where the only difference is the β estimates. In this model, β estimates are not fixed through the cutoff j but are allowed to vary. That implies a significantly higher number of coefficients to be estimated.

The relaxation of the proportional odds assumption might reveal asymmetrical effects (Fullerton & Dixon 2010). That means that the effect of variable changes across the individual cumulative logits. For example, the dependent variable might be composed of three ratings: low, medium and high. A variable X might positively contribute to the movement of people from the low rating to the medium|high, but might not contribute to their movement from low|medium to high. Fullerton & Dixon (2010) describe in detail why this phenomenon might happen and give a number of examples. Ordered Logit would fail to reveal these asymmetrical effects.

As a consequence of the lack of research on the topic of this thesis, there is no reason to assume the proportional odds assumption. Nevertheless, the results might later suggest the assumption to be fulfilled.

Model specification

This part gives an overview of the two main topics to be investigated. This is a summary of the hypothesised models, which are subject to modifications

Table 3.12: Abbreviations of variables in the models

Variable	**Abbreviation**
Password-Username Similarity	PUS
Password-Password Similarity	PPS
Password length	PassLen
The Effort	Effort
Gender	Sex
Literacy rate	Edu
Democracy index	Demo
Cybersecurity index	Cyber
Internet coverage	Internet
Mobile usage	Mobile
Polarity	Polarity

based on empirical observations.

First, Table 3.12 informs about the abbreviations of the variables used for both model families.

Model family 1: Password-Username similarity The Password-Username similarity was investigated under the Model family one label. The hypothesised relationships could be described by equation 3.6. It was believed the password-username similarity could be explained by password length, effort, sex, education, democracy level, cybersecurity level, internet coverage and mobile usage. β_j indicates the estimated effect of j-th variable, and ϵ is the error term. The equation follows the notation described in the previous part.

$$\begin{aligned} PUS_i = \beta_0 &+ \beta_1 PassLen_i \\ &+ \beta_2 Effort_i \\ &+ \beta_3 Gender_i \\ &+ \beta_4 Edu_i \\ &+ \beta_5 Demo_i \\ &+ \beta_6 Cyber_i \\ &+ \beta_7 IntCov_i \\ &+ \beta_8 Mob_i + \epsilon_i \end{aligned} \tag{3.6}$$

Model family 2: Password-Password similarity The Password-Password similarity (PPS) was investigated under the Model family 2 label. It was believed, the PPS could be explained by equation 3.7. It was expected that the

Table 3.13: Identified variables and the proxies per model

Category	Variable	Abbreviation	Model 1	Model 2
2*Password derived	Password length	PassLen	x	
	The Effort	Effort	x	
Microeconomic	Gender	Sex	x	x
5*Macroeconomic	Literacy level	Edu	x	x
	Democracy level	Demo	x	x
	Cybersecurity index	Cyber	x	x
	Internet coverage	Internet	x	x
	Mobile usage	Mobile	x	x
2*Dependent variable	Password-Username similarity		x	
	Password-Password similarity			x
Sentiment	Polarity		x	

PPS could depend on sex, education, democracy level, cybersecurity level, internet coverage and mobile usage. β_j indicates the estimated effect of $j-th$ variable, and ϵ is the error term.

$$\begin{aligned} PUS_i = \beta_0 &+ \beta_3 Gender_i \\ &+ \beta_4 Edu_i \\ &+ \beta_5 Demo_i \\ &+ \beta_6 Cyber_i \\ &+ \beta_7 IntCov_i \\ &+ \beta_8 Mob_i + \epsilon_i \end{aligned} \tag{3.7}$$

Table 3.13 gives an overview of the hypothesised models and the explaining variables. The following section describes the whole process from text-linguistic processing to the polarity estimation.

3.4 Sentiment

3.4.1 The notation

The goal of the Polarity analysis was first, to estimate the polarity of passwords and, second, its impact on password management. In any case, the polarity estimation was not an easy process. The password had to be split into meaningful words, which required models to split concatenated words into separated versions. On top of these segments, a polarity model was estimated. This section describes these steps.

Denote P as a password which is a sequence of letters, numbers and special characters. The password is composed of word segments and gap segments. Word segments are parts of a password that match precisely with words from a dictionary, and a gap segment is a sequence of special characters or extra letters with no special meaning.

Each password can be composed of multiple words and gap segments, and the goal is to find the correct segment combination. A *Split* is one possible way how to break a password into word and gap segments. The goal is to find the most probable split one can get. Table 3.14 demonstrate this terminology on two examples.

Table 3.14: A demonstration of the password segment notation

Category	Example 1	Example 2
Password (P)	helikescats123	123fish&chips
Word segment	he, likes, like, cat, cats	fish, chips
Gap segment	123	123, &
Password Split (PS)	e.g.: he like s cat s 123	e.g.: 123 fish & chips
The best Password Split	he likes cats 123	123 fish & chips

3.4.2 Word break problem

Under normal circumstances, polarity estimation is done using a set of labelled phrases. The model is then trained to estimate the polarity based on the words used and their order. The model might predict positive, neutral or negative connotations on the provided phrase.

Unfortunately, the data used in this thesis are passwords. That is a concatenated string (e.g., "helikesicecream"). Because of that, the very first step in the process is transforming raw passwords into an understandable version. That is, split the password into all possible segment combinations (i.e., the word and gap segment) and, among them, find the Best Password Split (i.e., the most probably split).

The first step was to find the segments. It was solved by an algorithm searching through the whole password and comparing sub-strings to a dictionary. The idea was to take a selected dictionary of words in a given language and generate all combinations of words that could be forming the password.

The idea behind the algorithm is following. Let $coverage[i]$ indicate possibilities how to split i first characters of a password into words using the selected dictionary D. Let $max\_coverage[i]$ be a number that represents how many characters up to ith character can be matched from the dictionary D. Let dictionary D be a vector of words from a language where every word appears exactly once.

For the demonstration of the algorithm, considers a password "dogsandcats". The algorithm tries to match the i first characters with the dictionary. If a match is found, we append this chunk to previously identified words. Next, a loop moving the start position until n-1 is performed. At this point, it is tested the chunk starting at the second position until ith character. Again, it is tried to match the chunk with the dictionary, and if successful, we append it to possible solutions. Subsequently, we increase the$i$ by one and rerun the inner loop.

An example can be found in Table 3.15. It is shown how the algorithm splits the password "dogsandcat". i indicates the outer loop increasing the chunk and ii indicates the inner loop stripping the chunk from left to the ith position. In the first round, the chunk "d" is evaluated. As it is not found in the dictionary, nothing is added to the solutions. At $i = 3$ and $ii = 1$, we evaluate the chunk "dog". It was found in the dictionary, and thus, it was added to the coverage[3], indicating that only the word "dog" was matched in the first three characters. $max_coverage[3]$ says that up to the 3rd character of the passwords, three consecutive characters were successfully matched with the dictionary.

When i reaches 4, we are evaluating "dogs". The word "dogs" is matched with the dictionary, and thus, it is appended to the solutions. However, as in the previous run, if we identified "dog" and "dogs" is not a prolongation of the previous word, we create a new branch of possible solutions resulting in ['dog', 'dogs'] as candidates for the solution.

The simple case is where two non-overlapping words are concatenated in the password. In this case, after identifying the first word, the second is appended to the solution. In case two words share characters (i.e., "dogs" and "sand" in a password "dogsand"), the word "sand" is not appended to the previous solution "dogs", but it is appended to the latest solution excluding the whole sequence "sand". It would be appended to the solution for $i = 3$ where "dog" was identified as a solution.

The algorithm is robust, and one can be sure it identifies all the solutions given a dictionary D. Nevertheless, as all words are tested at all positions, it

Table 3.15: Demonstration of the splitting algorithm (WBA)

i	ii	chunk	matched	max_coverage[i]	coverage[i]
1	1	d		0	[]
2	1	do		0	[]
2	2	o		0	[]
3	1	dog	dog	3	[’dog’]
3	2	og		3	[’dog]
3	3	g		3	[’dog’]
4	1	dogs	dogs	4	[’dog’],[’dogs]
...					
5	1	dogsa		4	[’dog’],[’dogs]
...					
6	1	dogsan		4	[’dog’],[’dogs]
...					
7	1	dogsand		4	[’dog’],[’dogs]
7	2	ogsand		4	[’dog’],[’dogs]
	...				
	4	sand	sand	5	[’dog’,’sand’],[’dogs’]
	5	and	and	5	[’dog’,’sand’],[’dogs’, ’and’], [’dog’, ’and’]
...					
10	1	dogsandcat		7	[’dog’,’sand’],[’dogs’, ’and’], [’dog’, ’and’]
...					
10	10	c			[’dog’, ’sand’, ’cat’], [’dogs’, ’and’, ’cat’], [’dog’, ’and’, ’cat’]

might eventually generate a vast number of solutions. This number is implied by the size and the granularity of the dictionary D.

Modification of the algorithm

The algorithm is as good as the quality of the dictionary D. Nevertheless, there will be a trade-off between the granularity (i.e., detail) of the dictionary and the number of possibilities the algorithm produces. The dictionaries used in this thesis were of high granularity, covering a large share of the potential words. That, unfortunately, increased computational requirements.

Consider the password

$$helikesicecreamanddogsandcatshelikesicecreamanddogsandcats$$

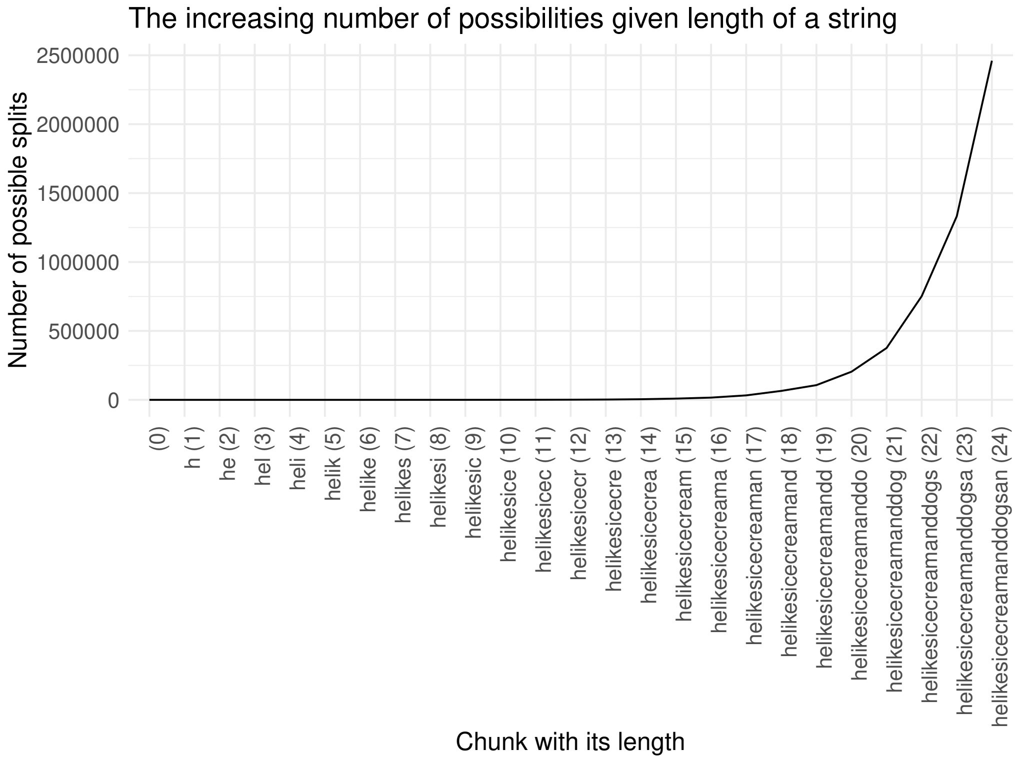
which is intended to break into

he likes icecream and dogs and cats he likes icecream and dogs and cats

This password was chosen because of its length, overlapping words (i.e., the chunk *dogsand* could be broken into the pairs $dogs, and$ and $dog, sand$) and because it contains prolongation of words (i.e., *dog* might be prolonged to *dogs*). Furthermore, the password is composed of short words (e.g., $dogs, cats, he$), which are harder to identify, especially in long passwords.

Figure 3.8 demonstrates how many possible splits were identified given n first characters of the subjected password. It can be observed that the number

Figure 3.8: Number of possible splits given first characters of the password



increases exponentially. It would be problematic to work with passwords longer than approximately 30 characters as the number of combinations explodes.

To deal with this limitation, a modified algorithm aiming to reduce the space of possible splits was introduced. The aim was to decrease the set of possible splits while retaining the true split in the set. For humans, that might be a trivial task (e.g., it might be evident that certain combinations do not make sense), while for the computer, it is not an easy task.

The assumption for the modified algorithm is that one should not have too many and too few words for a password. Too many splits would mean that the password is broken into pairs of words, and on the other hand, too few splits would indicate concatenation of words.

During every iteration, the modified algorithm decides which splits to keep and which should be discarded. This decision is based on several potential splits of the first n characters of the passwords.

Consider the same password as before. When the algorithm iterates and evaluates the sets of possible splits up to position n (e.g., for $n = 4$ we have [['dogs'], ['dog','s']]), the number of chunks in every option is evaluated. If the number of selected words is too high, it might be expected that it is artificially created as the words are shorter than expected (on average). The threshold was defined as the following:

$$Threshold = \frac{length\ of\ the\ chunk_i}{Expected\ length\ of\ a\ word} \approx number\ of\ breaks$$

Thus, using the expected length of a word, it is estimated how many breaks the chunk should have. If the number is higher than expected (i.e., words are shorter than expected), the given split is discarded. This approach is sensitive to the expected length of a word. Because of that, several thresholds were tested. This comparison can be seen in Figure 3.9. On x-axis lays a sample chunk with its length, and on the y-axis, one can see the number of different splits the algorithm would produce, given the threshold variable.

It can be observed that a threshold of 1.5 would not help much as it explodes too soon. The threshold of 2.5 delivers better filtering; however, it starts to increase dramatically after the 25th character. Keeping in mind that the lower the threshold, the safer, it was introduced incremental filtering. The threshold is rising with the string's length and is defined per intervals as described in Table 3.16.

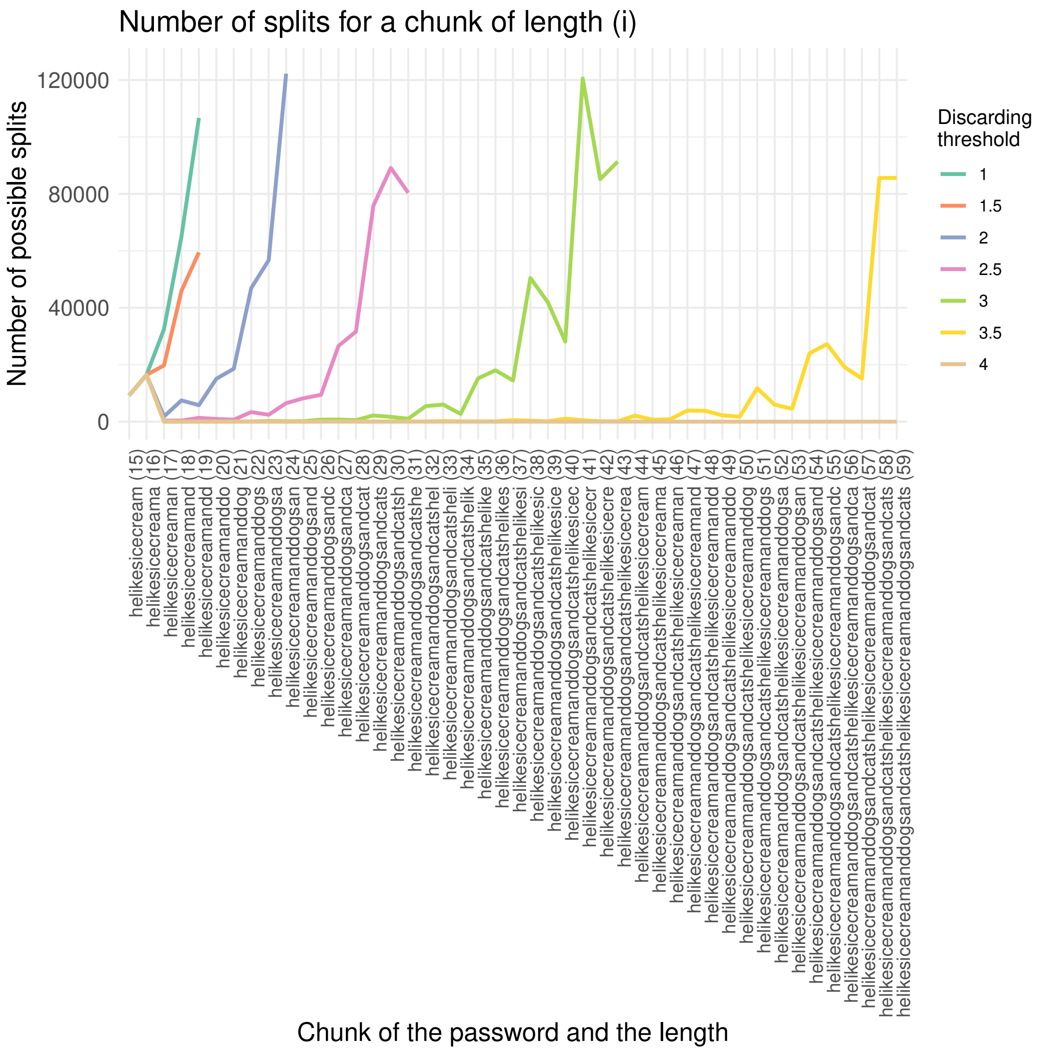
Table 3.16: Discarding threshold for splitting passwords

Length of a chunk	Threshold for filtering
1-15	None
15-20	2.5
20-30	3
30-40	3.5
40-45	4
45-50	4.5
50+	5

The filtering might eventually drop the real value. In order to make the filtering approach softer, an escape option in the loop was introduced. If the filtering is too restrictive and results in an empty set, the threshold is relaxed to 2.5 for that particular iteration. That assures that candidates will be maintained, while it increases the probability of retaining the real split.

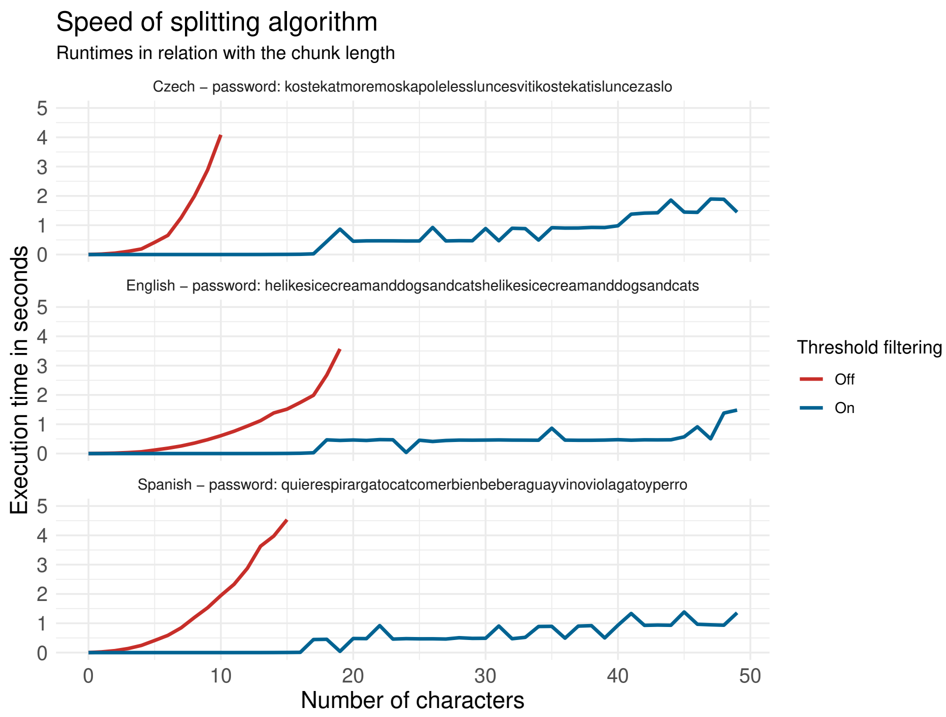
Figure 3.10 demonstrates the impact of the filtering approach on the speed of splitting the password. Czech, English and Spanish were chosen for this

Figure 3.9: Identification of discarding threshold



comparison. English is an example of a simple language, Spanish is a language with medium complexity and Czech due to its high complexity and high usage of diacritic marks.

Figure 3.10: Estimating the effect of filtering



It can be seen that without filtering, the number of possible splits explodes around the 20th character. An algorithm with such a performance would not be applicable as a password in the data set can be longer than 60 characters. On the contrary, the filtering reduces the execution time dramatically and makes it feasible to split a password made of 50 characters decently fast. For the Czech language, the algorithm on the complex password took around a second and a half. That is still relatively slow, but if only a few passwords of such length exists, the computation would be feasible.

In summary, the filtering method had to be applied as the raw algorithm produced too many unrealistic splits. Filtering is applied, and to mitigate the risk of discarding the real split, the dynamic threshold is set. Furthermore, no filtering is applied for chunks smaller than 15 characters. Finally, it should be reminded that an average password has around eight characters and thus, the vast majority of passwords will not be affected by the filtering at all.

Full code for the modified Word Break Algorithm can be found in Appendix in Codes in Listings A.1.

The dictionary D

The WBA algorithm requires a carefully selected dictionary of words. The dictionary is used to find potential word combinations forming the password. A description of such a dictionary is presented in the following section.

The leading dictionary for all languages consisted of the GNU Aspell[9]. This project is an automatic spelling checker currently supporting over 70 languages, easily accessible on any Linux bash. Despite the long history of the program, the latest version of the dictionaries come from October 2019. Thus, it was expected that the dictionary captured the contemporary language matching password's history.

The dictionaries were downloaded for the languages in the sample and exported as a text file. A few processing steps needed to be applied. First, all words were converted to lowercase. Passwords were also normalised by lowercase conversion. It would be interesting not to apply the lowercase normalisation, but users might intentionally change letters from lowercase to uppercase to improve the quality of the password. However, it is not feasible to quickly identify whether an uppercase letter in a password is intentionally changed or it is the feature of the word. Thus, for the polarity analysis, passwords and dictionaries were converted to the lowercase form without losing much information.

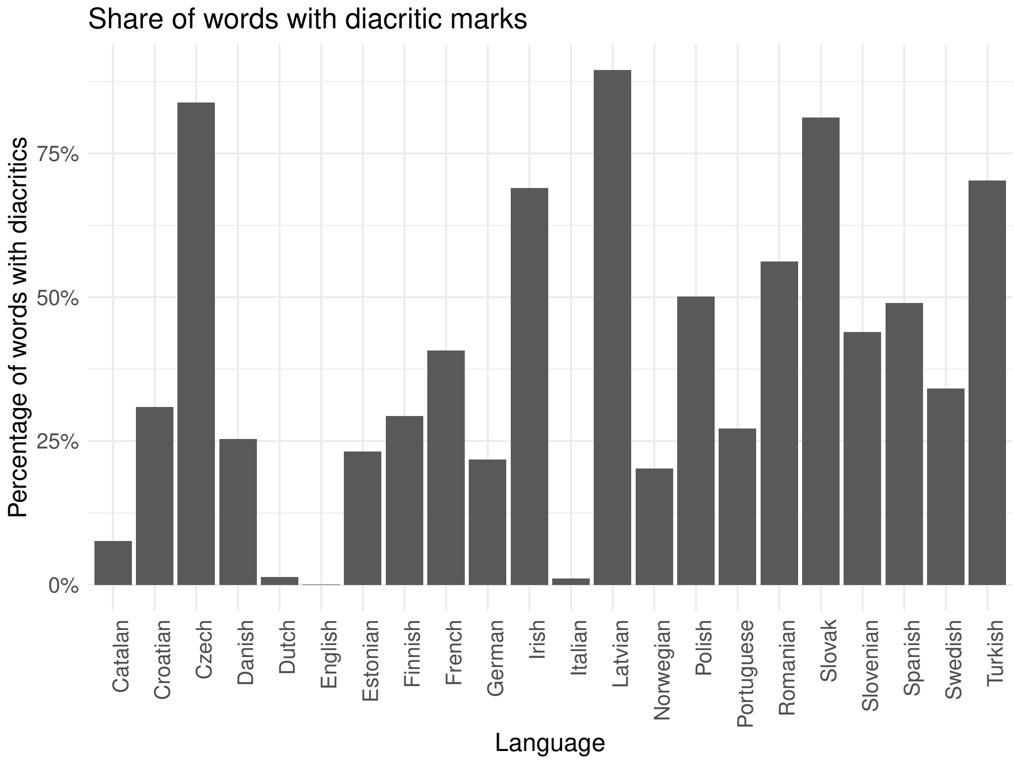
Next, in order to decrease the noise in the data, numbers were omitted. It was expected that numbers might not indicate the polarity of a password. Additionally, it was necessary to decrease the WBA algorithm's search space as much as possible as the size of the dictionary D negatively affects the speed of splitting. Nevertheless, numbers were considered for the use of age identification.

Last, a few languages use diacritic marks. In general, passwords can be composed of simple upper and lowercase letters, numbers and special symbols. Because of that, words with diacritics from the dictionaries were converted to their counterparts without diacritics. For instance, the letter "á" from the Czech language was translated as a. This might, unfortunately, introduce some noise to the data. Consider two words that are identical after the diacritic correction. In this case, it is not feasible to estimate the correct form of the word in the password. Luckily, not all languages use diacritics.

Table 3.11 reveals the percentage of words in the dictionary, including some

[9]http://aspell.net/

Figure 3.11: Presence of diacritics among languages

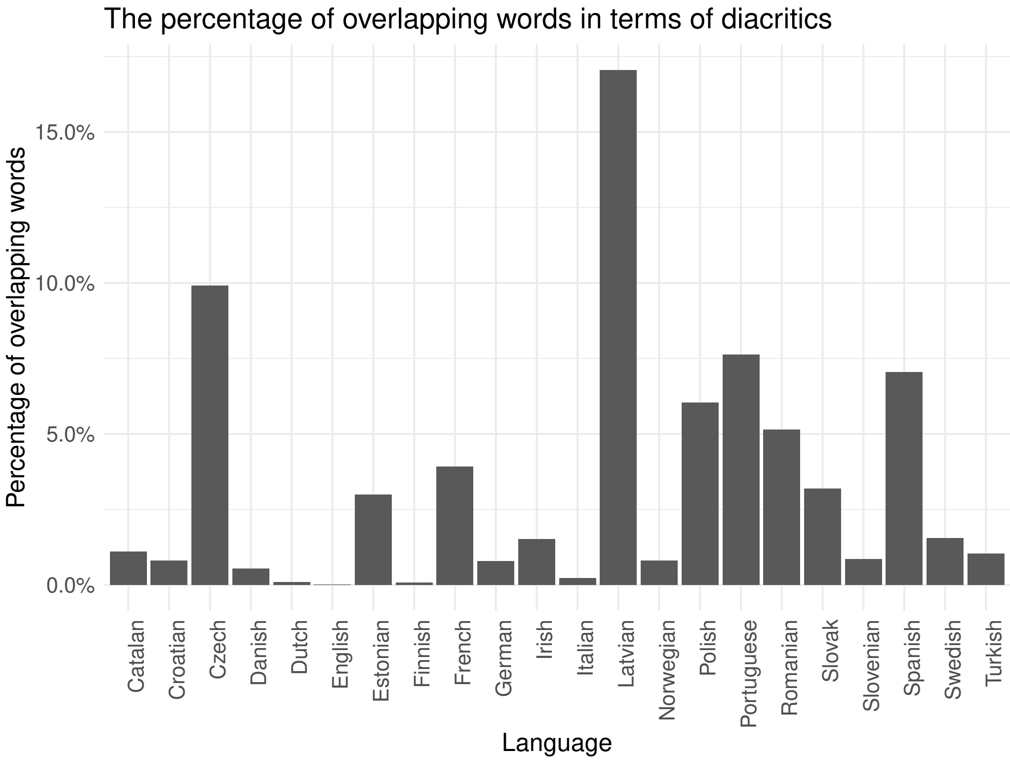


diacritics. A few languages, such as Czech, Slovak, Latvian or Turkish, demonstrate a high share of words with diacritics. However, the presence itself does not imply ambiguous translation. For example, if, for a given word with omitted diacritics exists exactly one counterpart in the original corpora, the translation is unequivocal and does not imply any bias. The Language Models are based on frequencies of n-grams, and they do not work with the meaning of the words. Thus, the Czech word "ještěrka" and the translated version "jesterka" would be treated in the same way. 3.12 indicates how many overlapping words existed in the dictionaries of selected languages.

The standardised Aspell dictionaries should contain a vast majority of words used by people. However, for further enhancement of the Word Break Algorithm results, modified Aspell was prepared. That consists of the original Aspell extended by a dictionary formed by words found in the language's raw corpora.

The corpora came from Lindat (discussed in detail in the following part). One hundred million rows were randomly chosen from the raw corpora as the search for unique words is computationally expensive, and the marginal gain of additional text is decreasing soon. Unique words from these corpora then

Figure 3.12: Overlapping words in dictionaries



enriched the Aspell dictionaries.

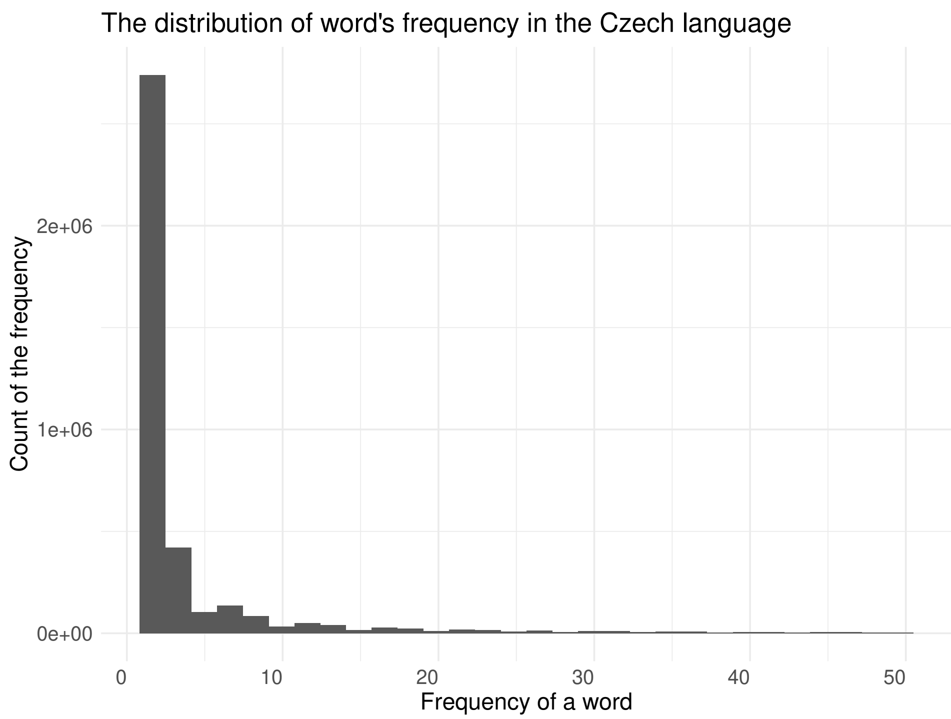
These modified Aspell dictionaries have their advantages and disadvantages. First, they are decently detailed, including mistakes that people generate. On the other hand, the high granularity implies the lower performance of passwords' parsing as more possible splits can be created. Furthermore, this is implied both by raw Aspells and corpora-based dictionaries; it might contain rare words to use. That means that the parsing function would generate a large number of possible splits that the Language Model would have to deal with.

It was attempted to deal with the noise induced by typos in data by applying simple filtering. The idea was to introduce a threshold of the frequency of occurrence, and a word would be included in the dictionary only if it appeared n times. Some languages are rich in vocabulary, and this approach did not deliver acceptable results. It was observed that the share of miss spelt words were lower than the share of rare words in a correct form. These observations suggested that the corpora included rare words not presented in the Aspell but being an existing word.

Further analysis was based on n-grams, taking into consideration the surroundings of a word. Thus, if this would be built on a sufficiently large corpus, the incorrect forms of words would be overpowered by the correct forms and no to small bias would be produced. The only drawback of leaving the wrong words in the dictionary is the longer running time, as the Word Break Algorithm would produce a larger number of possible splits given the larger corpora. The WBA is not aware of the popularity of the word in the language.

In conclusion, the presence of incorrect words in the dictionary passed to WBA should not affect the Language Models described in the following section, but it has a negative effect on estimation time.

Figure 3.13: Word frequencies



3.4.3 Language Models

Next task was to identify the optimal split out of all the combinations produced by the Work Break Algorithm using the dictionary. Language Models were trained and used for finding the best split among those combinations. The following section describe the training process as well as the approach in general.

The proper division of passwords was unknown. Thus, one option might be to guess it using a statistical approach. The WBA produced a large number of possible versions as it was based on extensive dictionaries of words. Thus, it was believed that in most cases, one of the versions should be the correct version.

Language Models can calculate the likelihood of occurrence of a sequence of words regarding provided corpora. As the Language Models were trained on large corpora, it was believed that they should choose the most likely password split decently well.

Formal notation of the n-gram Language Model N-gram Language Models are derived from Markov Chain approximation. The probability of the current word is determined by n previous words. Let $W = w_1, w_2, \ldots, w_i$ be a sequence of words. Equation 3.8 indicates a general idea of the n-gram application. The equation determines the joint probability of the sequence. Equation 3.8 is also a simple version of the n-gram Language Model.

$$p(W) = \prod_{i=1,\ldots,d} p(W_i | W_{i-n+1}, W_{i-n+2}, \ldots, W_{i-1}) \tag{3.8}$$

Consider the phrase *I like cats and dogs.* The probability of this phrase using tri-grams would be following:

$$\begin{aligned} P(\text{"} < s > i\ like\ cats\ and\ dogs\ \ < e > \text{"}) = {} & P(i \mid\ < start >,\ )\cdot \\ & P(like \mid i,\ \ < start >)\cdot \\ & P(cats \mid like,\ i)\cdot \\ & P(and \mid cats,\ like)\cdot \\ & P(dogs \mid and,\ cats) \end{aligned} \tag{3.9}$$

KenLM implementation KenLM is an efficient implementation of the n-gram based Language Models by Kenneth Heafield (Heafield 2011). This implementation is fast, memory-efficient, capable of using multi-core processors and is open source. Enables usage of efficient data structures (e.g., trie) and pruning for space and speed optimisation. Furthermore, it is capable of handling gigabytes of data on a conventional notebook. The library was compiled on Linux.

Author claims the package is faster and less memory demanding then a variety of existing solutions (e.g., IRSTLM [10] BerkeleyLM [11] or SRILM [12]).

[10]https://hlt-mt.fbk.eu/technologies/irstlm
[11]https://code.google.com/archive/p/berkeleylm/
[12]http://www.speech.sri.com/projects/srilm/

Training of the models The models were trained on large corpora consisting of a general language. The corpora came from the CoNLL 2017 Shared Task - Automatically Annotated Raw Texts and Word Embeddings [13]. This is a collection of annotated text by the Universal Dependencies project[14]. The annotations themselves were not useful for this thesis, but the files contained large raw text that could be extracted. In total, there were 5.9 billion sentences occupying 630GB on the drive.

The corpora and passwords should be processed similarly. As it was aimed to split the textual part of a password into words, numbers and special characters were removed from the corpora. Furthermore, if applicable, diacritics were also removed from the corpora, which corresponds to the preprocessing of dictionaries used for WBA.

As for the KenLM model, it was necessary to optimise the order of the underlying n-grams. A common choice is 3 to 5 order n-grams. However, for the higher-order n-grams, extensive data are required (Jurafsky & Martin 2019). In this thesis, the targeted text differed from a normal text significantly. Most of the passwords are collections of one or two words. Rarely, it was a composition of three or more words. Thus, it was expected that lower-order n-grams would perform equally well to higher-order ones.

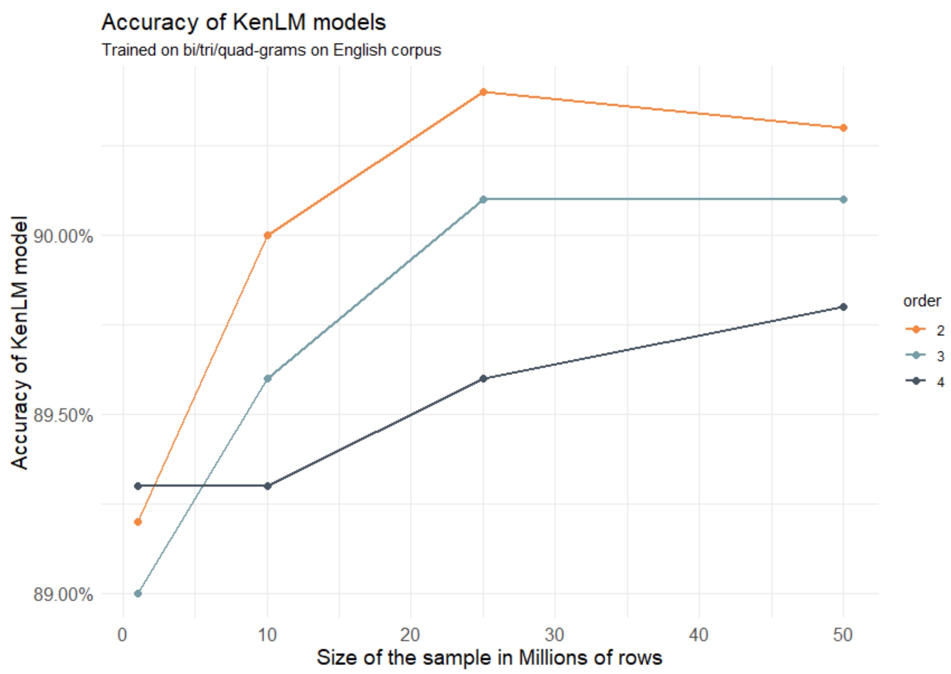
In addition to the order, it was necessary to deal with the corpora size. They had up to 30+ gigabytes per file. It was expected that at some point, the marginal effect of additional data on the performance of the Language Model would be negligible. Several models were trained on different size of the corpora and different orders of the n-grams. Then, for each language, sentences were concatenated and then split with the trained models. Accuracy was computed as the percentage of correctly split phrases. Figure 3.14 reveals the accuracy of the models for the English language.

As one can see, the accuracy increases with the size of the corpus. However, it culminates at approximately 25 millions rows of the text data. Furthermore, the accuracy is only slightly affected by the order of the n-gram. The highest accuracy is achieved by the Language Model built on bi-grams. This is caused by the pruning of the model, which drops too rare observations.

For the rest of the languages, the data looked similarly. Thus, it was decided that the optimal size of the corpora would be 25 millions rows. That is an

[13] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1989
[14] http://ufal.mff.cuni.cz/udpipe

Figure 3.14: Identification of n-gram order and corpus size



amount combining enough information (occupied around 10GB per language on the drive) and acceptable performance of the models.

Specification of the corpora per top-level domain The Language Models were not estimated for all domains. The availability of corpora implied this restriction. As the corpora source was Lindat mentioned before, only languages presented here were taken into consideration to ensure consistency in the analysis. The inclusion of all relevant languages is one of the suggested improvements as it was not feasible to cover everything in this thesis. In total, 22 languages were identified.

On one hand, there were the corpora languages, and on the other hand, there were Top Level Domains. Now, the task was to assign correct languages to TLDs. A list of official languages was used as a primary source. Furthermore, if there was a significant non-official language, it was included as well. All training sets of official languages were extended by English corpus as it is widely spoken and understood as an international language. In total, 22 languages were identified as feasible. That corresponds to 140 top-level domains. Table A.9 reveals what languages were considered per domain and the list of domains itself.

At this point, passwords were broken into expected words by the Language

Models. The next step was to develop a methodology for the polarity assessment.

3.4.4 Sentiment methodology and data

The difficult aspect of the sentiment analysis of passwords was first, the general shortness of passwords (an average password had around eight characters) and second, the number of languages in the sample.

Initially, it was considered to use polarity dictionaries (i.e., dictionaries of words with annotated polarity) as it would allow considering a large number of languages in the analysis. There are large polarity dictionaries including dozens of languages (Chen & Skiena 2014).

Nevertheless, it was realised that while it would be possible to study nearly all languages in the sample, the results might be significantly misleading. Consider the word *terrible*. The polarity of the word in the phrase "The movie was terrible" is negative. However, in the phrase "The meal was terribly delicious", the polarity should be positive. Moreover, a basic model based on polarity dictionaries would fail to distinguish these two forms.

Model selection

There are a few different approaches on how to approach Sentiment Analysis based on annotated text. The standard way using Baesyan statistics (Suppala & Rao 2019) or the modern approach using Neural Networks (Chen *et al.* 2017). These models are trained on labelled data. That is, on a collection of documents where each document has a label indicating its sentiment. Most frequently, the label is "Positive" or "Negative". Some collections also include "Neutral" as an additional label.

For the Sentiment Analysis based approach, it was necessary to find labelled data for multiple languages. The model would be as good as the training data match with the tested data regarding the structure and type. Since passwords are generally concise texts, the most frequently used data like books would not approximate the password data correctly. Standard text is much longer than passwords, and authors have entirely different targets when writing a book and setting up a password. Therefore, it was essential to find a text with a similar length and structure.

Analysis of Twitter Sentiment is notoriously known among NLP researchers and frequently used for Sentiment Analysis (Shelar & Huang 2018; Suppala & Rao 2019; A. & Sonawane 2016; Martínez-Cámara *et al.* 2014).

Tweets are frequently very short. By default, there is a limitation to 240 characters, and the average length of a tweet is only 28 characters 28[15]. Furthermore, due to the length limitation, one cannot consider the text as coherent or fluent - users use abbreviations and short phrases to meet the limit.

Because of that, Tweets could be a good approximation of the password language. However, it was necessary to admit that it was not the perfect approximation, as passwords are even shorter and frequently do not even form a phrase. Nevertheless, it was a decent compromise of data availability and similarity with the password language.

Twitter data has the advantage of being used by various nations, and thus, data could be found in various languages. That was more than welcome as the same source and structure of the data should contribute to the consistency of the Sentiment Analysis models.

Developing the polarity model First limitation in terms of languages was set by the Language Models described in the Language Models section. Thus, there were 23 candidate languages for Sentiment Analysis. While several papers were focusing on Twitter Sentiment Analysis, the availability of labelled data was sparse. Furthermore, the assessment of the sentiment was related to the annotator, and thus, it was desirable to use data from a recognised institution and data that were labelled under the same standards.

Slovenian researchers published an extensive study on 15 European languages (Mozetič *et al.* 2016). They put together over 1.6 million annotated datasets and made them available through the Clarin project [16]. The data contained the tweets' id and the annotated sentiment (i.e., positive, negative or neutral). The tweet itself was missing, and thus, the raw tweets had to be downloaded through a custom python script using the Twitter API and a developer account.

While authors offered 15 languages, not all of them matched with the selection from the Language Models part. Languages such as Albanian and Bosnian were not used for Language Modelling as Lindat did not publish the corpora.

The intersection of Language Models and the Twitter data resulted in 9

[15]https://smk.co/article/the-average-tweet-length-is-28-characters-long-and-other-interesting-facts

[16]https://www.clarin.si/repository/xmlui/handle/11356/1054

languages that could be used for the Sentiment (Polarity) Analysis. Namely Croatian, English, German, Polish, Portuguese, Slovak, Slovenian, Spanish and Swedish. That was a significant reduction in the number of languages. The availability of annotated Twitter data imposed notable limitations to this thesis. On the other hand, opting for consistency rather than quantity should imply higher quality of the results.

Description of the Twitter data

Figure 3.16 reveals the number of successfully retrieved Tweets per language. Altogether, there were 701 926 retrieved Tweets, which was significantly less than 1.6 million claimed by the authors (Mozetič *et al.* 2016). Unfortunately, some labelled Tweets by the authors were no longer available to the public.

There were more than 150 thousand labelled Spanish Tweets and less than 50 thousands of Tweets for Swedish, Spanish being the most frequent language in the dataset and Swedish being the least frequent one.

Figure 3.15: Number of retrieved Tweets per language

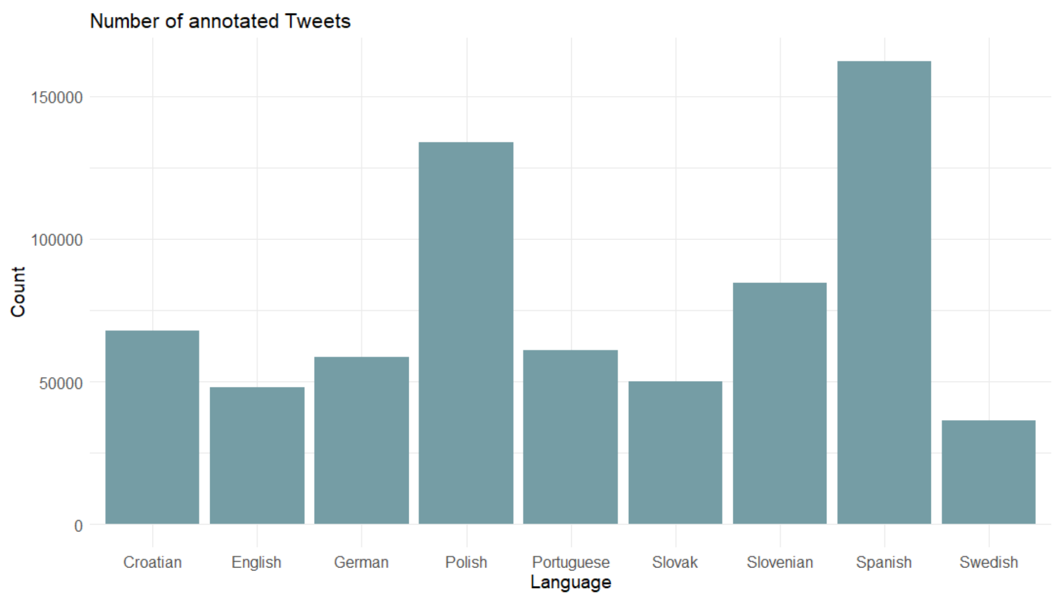
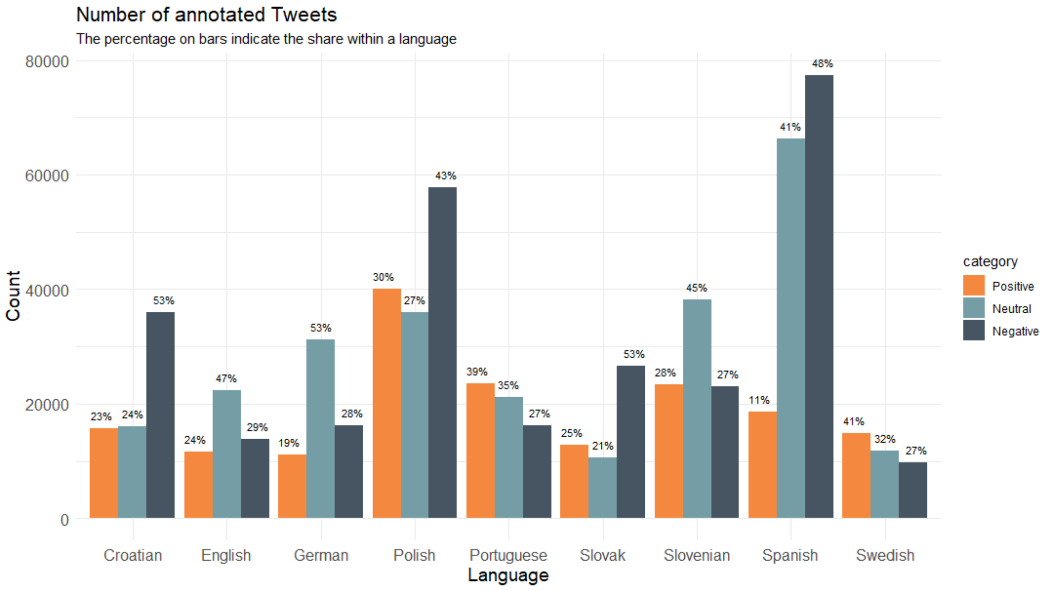


Figure 3.15 reveals the distribution of the labeled Sentiment in the data per language. In the majority of languages, there were no extreme inequalities among the polarity occurrence. Nevertheless, in Spanish Tweets, Positive tweets were relatively sparse, accounting for about 11% of the Spanish Tweets.

Figure 3.16: Number of Sentiment type of Tweets per language



Model selection

As mentioned before, the Bayesian approach was chosen to model the sentiment due to its decent implementation and promising performance.

The approach was based on calculating a large matrix indicating the presence of a specific word in the Tweet. This matrix was based on training data vocabulary with selected optimisation, such as eliminating too rare words. That contributed to a decrease in the noise and improved accuracy and time estimation at the same time.

Pure matrix of word occurrence would be too naive, and capturing polarity would be imprecise. That is due mainly to the changing meaning of words depending on the context. To help capture the context, the word occurrence matrix was extended by an n-gram occurrence flag indicating the word's context.

However, as Tweets are rather short, and the passwords are very short, n-grams should only help with relatively small n (e.g., bi-grams or tri-grams), as there would not be enough words to use higher-order n-grams, such as octa-grams. Table 3.17 demonstrates such a matrix. This approach is also called a Bag of Words.

Similarly to the WBA dictionary and corpora used for Language Models, tweets were preprocessed similarly. The list of the steps is the following:

1. Use TweetTokenizer from NLTK to eliminate redundant elements

Table 3.17: Demonstration of the Sentiment data structure

Text chunk	he	likes	cat	he likes	likes cat	you	likes you	icecream
He likes cat	1	1	1	1	1	0	0	0
He likes you	1	1	0	1	0	1	1	0
Icecream	0	0	0	0	0	0	0	1

2. Transform Tweets into lowercase

3. Omit hyperlinks

4. Eliminate special characters

5. Drop numbers

The *TweetTokenizer* from NLTK library[17] helps to drop redundant elements from a Tweet. Namely removes Twitter username structure from the text and limit the repetition of a letter in a word (i.e., *hiiiii* to *hi*).

The matrix was created using the TfidfVectorizer from the sklearn package[18]. This function allows for parameters modifying the resulting matrix. Relevant parameters investigates for the Sentiment models were *n-gram range*, *max features* and *stop words*.

n-gram range helps specify the length of n-grams used in the matrix. It is specified as a tuple (i,j) indicating the range that will be applied. For example, setting *i* to 1 and j to 3, n-grams up to order three would be applied.

max features controls the size of the matrix. The increasing number of included words raises the resulting matrix that could result in a matrix too big to be processed in case of large training data. At the same time, one might drop words that occur only once, as they might be typos or artificial words. This parameter controls the number of resulting features through term frequency.

stop words are words that do not bear a relevant meaning to the modelling. They might appear in the majority of documents and do not improve the performance of a model. These words are languages specific, for example, in the English language, *I, me, we, our* or *its* are considered to be stop words and are omitted from modelling.

The modelling was done in two major steps:

1. First, identify models that describe well the polarity per a single language

[17] https://www.nltk.org/api/nltk.tokenize.html
[18] https://scikit-learn.org/

2. Second, combine languages depending on the languages spoken in a country and train the final model

The first group of models was not used for the estimation of the polarity of passwords. Their purpose was to estimate how the model might eventually perform. The second group of models was based on multiple languages. It was expected that their performance would be worse than the first group's performance because the combination of languages is harder to model. Thus, the models on a single language served as a benchmark to measure the performance.

Language specific Sentiment models

The preprocessing for both groups of models were done identically. This task was treated as a Machine Learning task. The model development was designed in a general way so it could be applied to all languages individually as well as on groups of them.

At every iteration, the dataset was divided into train and test part. The test part was set to be 15% of the sample and used for the final estimation of the performance. Optimal parameters of the model (if necessary) were found using a grid search over a defined set of parameters. Four fold cross-validation were applied for improved performance estimation. For measuring the accuracy, F-Score, Recall and Precision were used to compare the models. These widely used metrics are defined as follows:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{3.10}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{3.11}$$

$$F\text{-}Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3.12}$$

True Positives stand for the number of correctly labelled positive observations by the model. *True Negatives* stands for correctly labelled negative observations. *False Positives* are observations that were labelled as positives but, in reality, are negative. Last, *False Negatives* are observations labelled as Negatives but, in reality, are Positives.

Precision indicates what share of positive predictions was actually correctly labelled. Recall reveals what proportion of actual positives were identified correctly. As one of the measures might be high and the other very low, F-Score helps trade-off between them. If either Precision or Recall is small, the F-Score is low. Both Precision and Recall have to be decently good to obtain a good F-Score.

Based on the literature and best practice, a few classic Machine Learning models were considered. Namely: Logistic Regression (Cox 1958), Naive Bayes (Minsky 1961), Support Vector Machines (Boser *et al.* 1992), Random Forest (Breiman 2001) and Decision Trees (Moore 1987). For these models, relevant hyperparameters were tuned using the mentioned grid search. Surprisingly, the best performing model was Logistic Regression, and it was best (in terms of the F-Score) for all languages in the sample. The performance of the other models, however, was only slightly worse. That was not a big surprise as the independent variables are a set of many binary indicators forming a significantly sparse matrix.

Figure 3.17 reveals the performance of the model per language. Overall, the F-Score oscillated around 0.6. That means the model was still much better than a random guess (considering three possible outcomes). The worst model was based on the Portuguese language falling below an F-Score of 0.5.

These models were estimated to see how the Logistic Regression models well the polarity in the selected languages. Furthermore, they set a benchmark for domain-specific models to estimate the expected domain-specific Polarity Models.

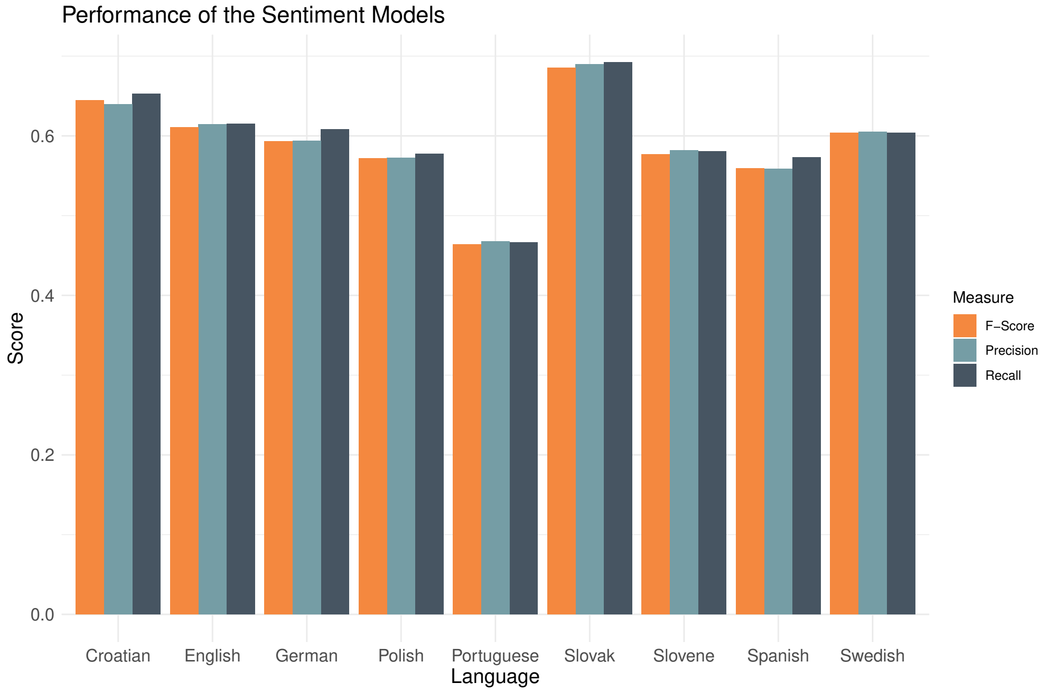
TLD specific Sentiment models

After estimating the benchmark models specific to a single language, it was necessary to adapt these models to Top Level Domains. That is, build models that would estimate the polarity per country or region. As discussed earlier, nine languages were included in the Polarity analysis.

The number of TLDs where the nine processed languages would cover all the spoken languages would be relatively small. Thus, in order to extend the number of TLDs for the analysis, a TLD was taken into account if at least one of the official languages was covered by the labelled twitter data. That is, at least one of the nine languages has to be an official language in a country.

Of course, the performance of a model in a country with five official languages where the Twitter data covered only one would not be high. On the other hand, it should only imply many neutral labels as the model does not

Figure 3.17: Performance of the baseline Sentiment models



know the other languages. Moreover, this ease of inclusion allowed us to work with dozens of domains.

The potential imprecision might be following. Consider the case of the Czech Republic. The official language is Czech, a significant share of the population speaks the Slovak language, and the Polish language is decently frequent. Unfortunately, the Czech tweets were not labelled by the Slovene Lindat. On the other hand, Polish and Slovak language was covered well. Thus, the sentiment for the Czech TLD would be trained on Slovak and Polish data. One of the first questions would be how the performance of such a model is affected and how the missing Czech data would affect the results.

In the hypotheses, it aims to identify the presence of sentiment and estimate whether it is positive or negative. If in a password occur words that were not seen in the training data (i.e., the labelled tweets), the model will return a constant value, which would be the Neutral label. If the password contains words seen in the train data, the polarity would be predicted according to the logistic regression weights.

Nevertheless, the critical point is that if the model does not contain Czech (and a significant number of passwords would contain Czech words), the large number of possibly mistaken Neutral labels would introduce noise to the data and decrease the significance of such variable.

If there would be many incorrectly labelled Neutral passwords and such a variable would be considered significant in the regression, that might suggest even more vital signs of the true polarity.

On the other hand, if the polarity turns out to be insignificant, there might be two explanations. First, it might be because the variable is genuinely insignificant or second, it might be mistakenly insignificant as the missing language that produced many Neutral labels overweight the polarity implied by the rest of the presented languages.

In addition to the official widely spoken languages in a country, the English language was appended to the models as it is a widely spoken language and frequently occur in the data. That might not be the case of countries, for example, under the Russian language's influence where the Russian is considered to be important. Nevertheless, such countries do not appear in the final sample of TLDs.

The approach for the model estimation was identical to the language-specific Sentiment models. That is a grid search over a set of potential models and hyperparameters with four-fold cross-validation. Similarly to the benchmark models, Logistic Regression was the best performing model, delivering the best F-Score among the considered Machine Learning models.

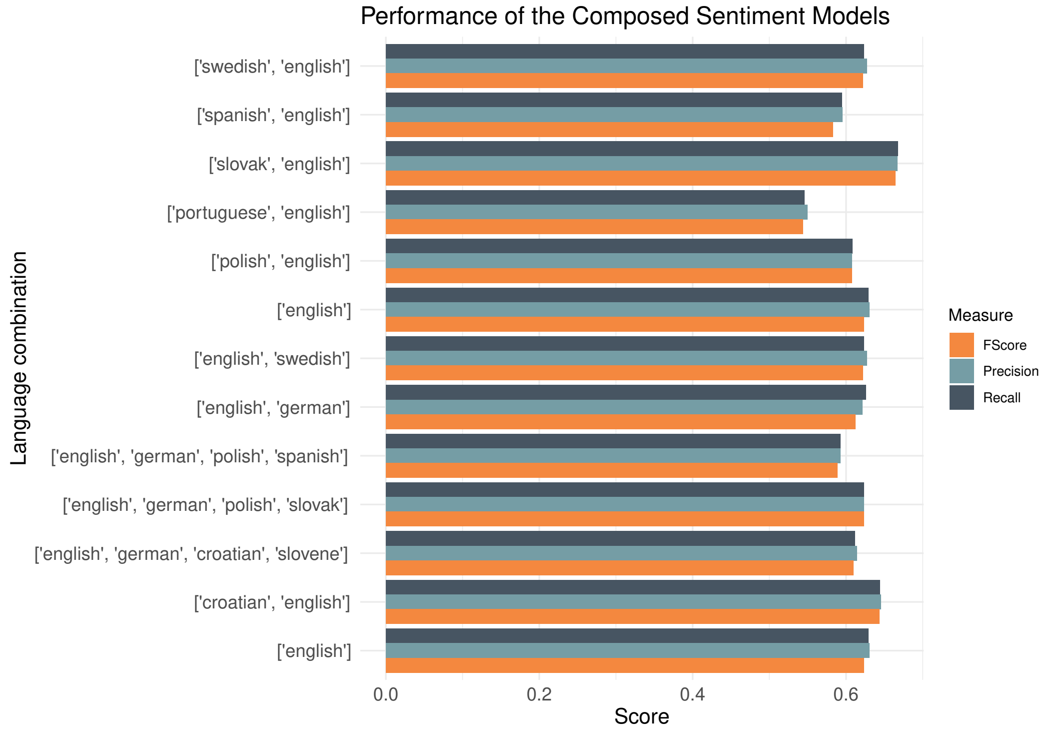
Figure 3.18 indicates the Polarity models' performance based on a combination of languages. Considering that three labels compose the target, all models were better than a random guess. A positive finding was that the spread between the Precision and the Recall is small in all the models. The worst performing model is based on Portuguese and English, having an F-Score of 0.54. On the other hand, the best performing model is based on the Slovak and English languages, achieving an F-Score of 0.66. In general, all the models perform similarly, having F-Score around 0.6.

The performance was only decent. Nevertheless, given the task and relatively straightforward approach, the results were acceptable.

Figure 3.19 reveals the proportion of the Positive and the Negative predicted labels per language. There were no expectations in terms of this distribution. One can see that overall, the proportion of identified sentiment was relatively small. It ranged from 0.25% to almost 3.5%. As expected, the Positive sentiment dominates the predictions. However, in 5 cases, the negative sentiment was more frequent than the positive one.

The lower share of other than Neutral Polarity might be explained by: a)

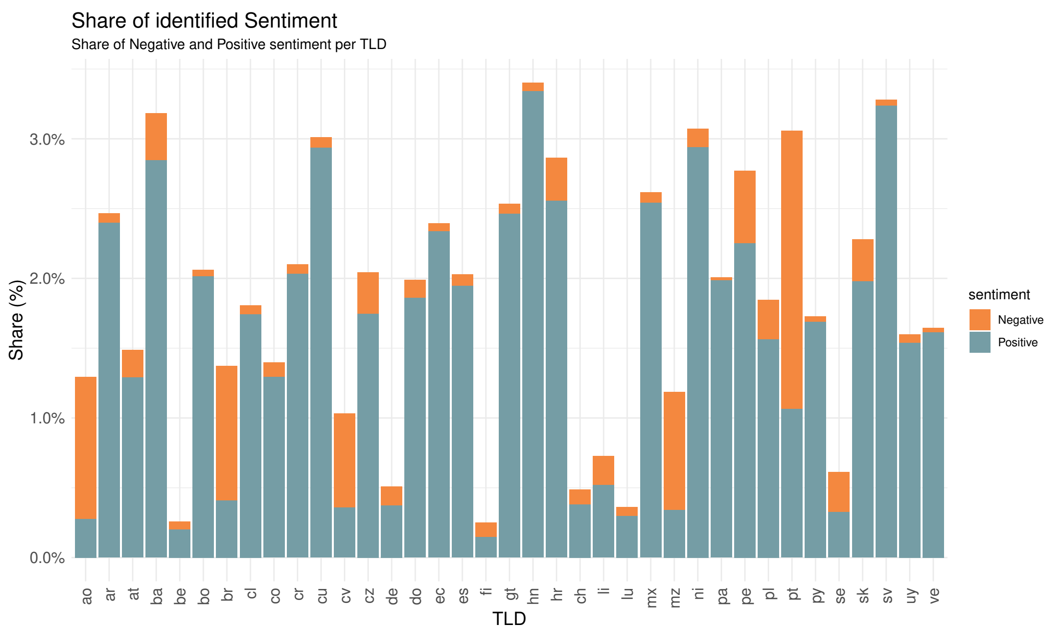
Figure 3.18: Performance of the Sentiment models



low tendency of users to employ positive and negative connotations to their password or b) inability to estimate the polarity based on Twitter data.

In conclusion, the identified sentiment is not negligible, behaves as expected and will be used in further calculations.

Figure 3.19: Share of a Positive and a Negative sentiment per TLD



3.5 Descriptive statistics

In this part, one can find descriptive statistics of the data prepared for Model 1 and Model 2. The data described in this part were derived as outlined in section Data.

3.5.1 Descriptive statistics of Model 1 data

This data sample contained information about the gender, password length, the Effort and Macroeconomic variables. Additionally, a further breakdown per country is presented as well.

In total, there were 395 623 362 relevant observations to this model. Table 3.18 indicates the target variable's statistics measured as the Levenshtein Distance between the username and the password and the password length. The PUS ranged from 0 to 35. 0 means that the two strings were identical, and 35 that hey they were very different. Surprisingly, the mean was 9.86, indicating that the username and password are not that similar on average.

The length of a password varies from 0 to 30. The average length is eight characters, which is not a high number. Eight characters long password is relatively weak.

Table 3.18: Sample statistics of data for H1 - Target, Length

Variable	**Min**	**Mean**	**Max**	**SD**
PUS (password-username similarity)	0.00	9.86	35.00	3.46
Password length	0.00	8.50	30.00	2.65

Table 3.19 reveals the sample statistics for Gender identification and The Effort. 15.4% of usernames were identified as belonging males and 12.5% as females. Around 70% of data were not linked to a given name. A positive finding was a similar share of decoded females and males.

Regarding the Effort, one can see that nearly half of the users used only one type of character set in their password. This number was alarming as these passwords are easy to guess (users use only lowercase letters, uppercase letters, numbers or special characters). Nearly 45% of users use at least two character sets in their password. 6% of users used three character sets, and only 0.4% of the most responsible users used four character sets. A lower share of passwords, including all character sets, was expected as more sets are harder to remember and write.

Table 3.19: Sample statistics of data for H1 - Gender, The Effort

Gender			
Males	*Females*	*F/M*	*Unknown*
15.4%	12.5%	4,3%	67.8%

The Effort			
Category 1	*Category 2*	*Category 3*	*Category 4*
49.0%	44.1%	6.1%	0.4%

Table 3.20 indicates the descriptive statistics for macroeconomic variables. The Democracy Index ranged from 1.52 to 9.80, with a mean of 6.18. The data seems to be bell-shaped, and countries seem to be rather democratic according to the index. Furthermore, data were available for 164 countries, making the variable to most sparse in the selection. There were around 20% of missing values.

Table 3.20: Descriptive statistics of Macroeconomic variables

Variable	**Min**	**Mean**	**Max**	**SD**
DemIndex	1.52	6.18	9.80	1.94
MobileCell	27.41	138.22	345.32	21.17
NetUsage	1.31	77.63	100.00	9.65
SecIndex	0.004	0.84	0.93	0.08
Literacy	22.31	98.89	100.00	2.72

The mobile phone possession ranges from 27% to enormous 345% of the population. That means that there are countries where mobiles are not common, and on the other hand, there are countries where an average person used three mobile phones. There are data for 177 countries which implies approximately 10% of missing data.

The Internet usage ranges from 1.3% to 100%. That means there were countries where the vast majority of the population did not have access to the Internet, and on the other hand, there were countries where all persons could access the Internet. The mean is 54%, indicating that the sample distribution should not be heavily skewed. There are less than 2% of missing data.

The Security Index ranges from 0.004 to 0.93. That indicates that there were countries with very poor cybersecurity, and on the other side of the spectrum, there were countries with excellent cybersecurity level. There were almost 10% of missing data.

Regarding the Literacy rates, the variable ranges from 22.3% to 100%. The average is 86% literate people. That means that there were countries with inferior ability to read and write. However, most of the countries should be rather educated.

Further breakdown by country It was expected that there might be significant country fixed effects. Therefore, a brief analysis of the country dimension was included.

Figure 3.20 indicates the distribution of the Literacy rate across countries. As one can see, the developed world is associated with high levels of literacy rate while countries in middle Africa shows significantly lower levels. Figure 3.21 reveals the distribution of the literacy rate. In line with the map, the vast majority of countries are highly literate. However, minimal literacy is around only 0.2.

Figure 3.20: Distribution of the Literacy rate across countries

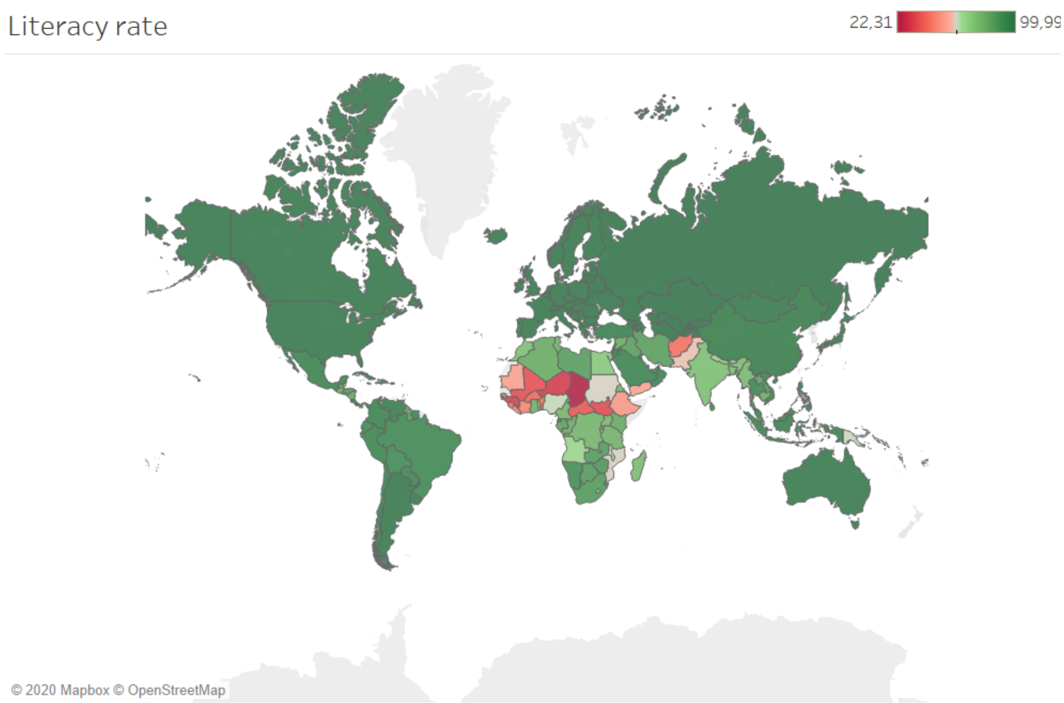


Figure 3.21: Distribution of the Literacy rate

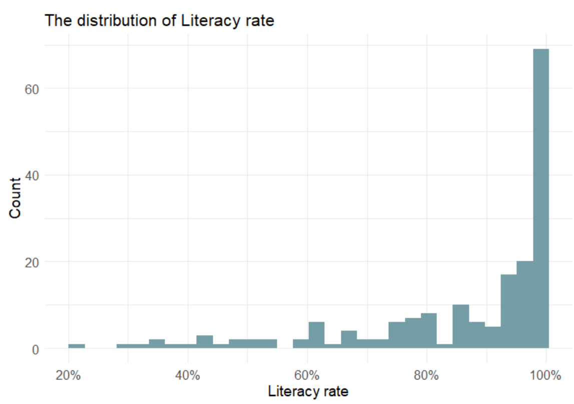


Figure 3.22 describes the distribution of the democracy index across countries. North America, South America and Europe are mostly democratic. On the other hand, most of the African countries and Asian countries are rated as less democratic. Figure 3.23 reveals the distribution of the variable. It seems to be neither uniform nor bell-shaped. Nevertheless, it seems that there are two groups of countries. One rather democratic with a mean of the index around 7, and the second group with less democratic countries with a mean around 3.

Figure 3.22: Distribution of the democracy level across countries

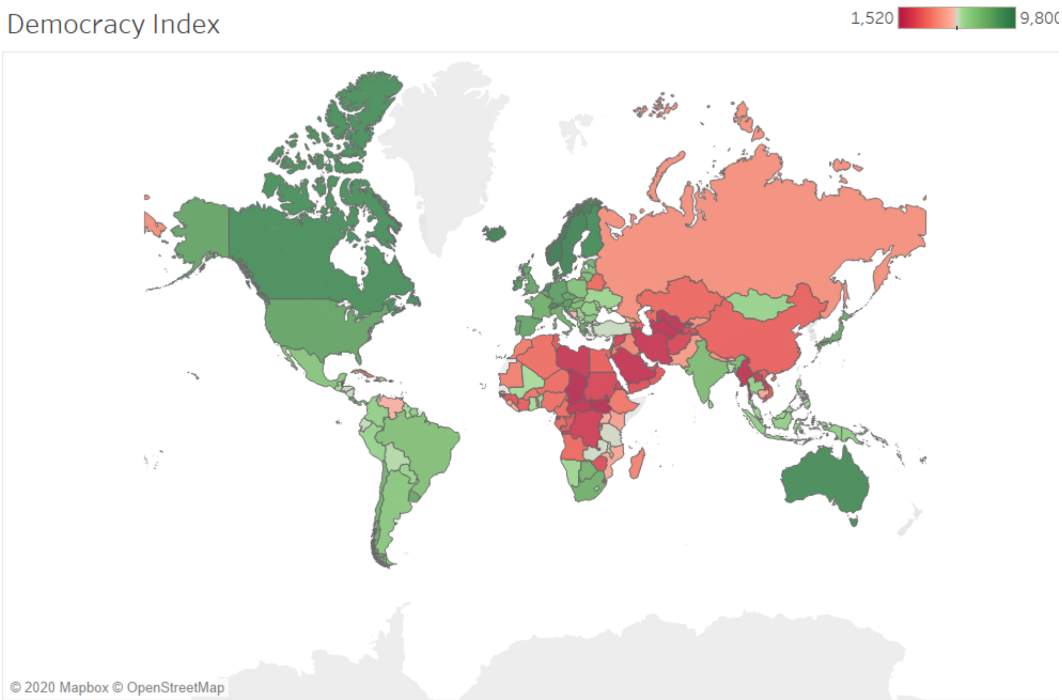


Figure 3.23: Distribution of the democracy level

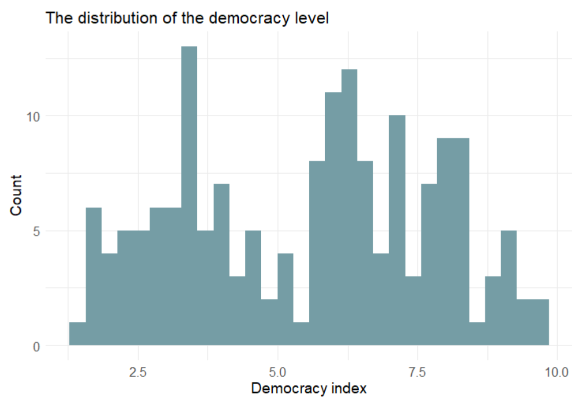


Figure 3.24 informs about the mobile usage distribution across counties. As discussed earlier, there were a few countries with extremely high usage. Up to 345% of people had a mobile phone. That is, the average person used more than three mobile phones. Leaving the countries with high usage aside, the variable seemed to be related to the country's wealth. America, Europe, and most Asian countries have a decent percentage of people with a phone, while poor African countries stayed well behind them, falling to 25% of people with a phone. Figure 3.25 reveals the distribution. It seems to be bell-shaped, with a few outliers in the upper part.

Figure 3.24: Distribution of the mobile usage across countries

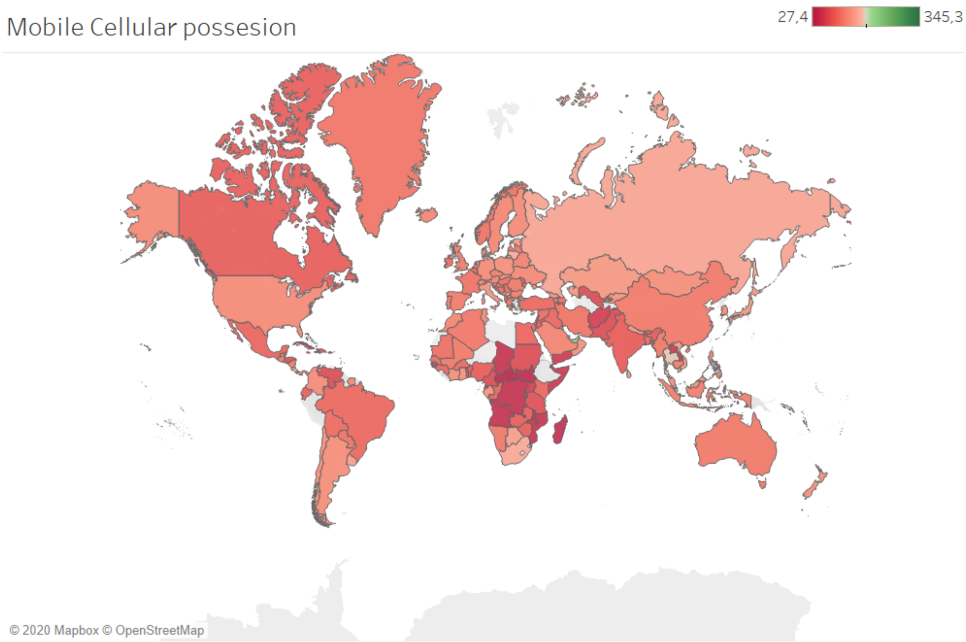


Figure 3.25: Distribution of the mobile usage

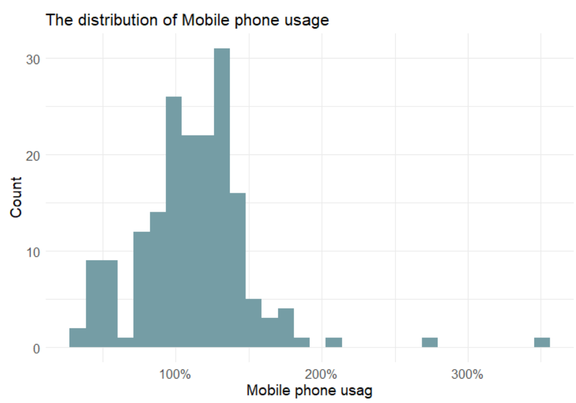


Figure 3.26 reveals the internet usage. As one can see, the developed countries demonstrate a high percentage of the population with access to the Internet. On the other hand, it seems to be harder for people in African countries to access the Internet. The figure can be as low as only 1.3% of the population. Figure 3.27 gives an overview about the distribution of the internet usage. Surprisingly, it was approximately uniform with a small break around the value of 40%. The distribution suggests the division into developed and developing countries.

Figure 3.26: Distribution of the internet usage across countries

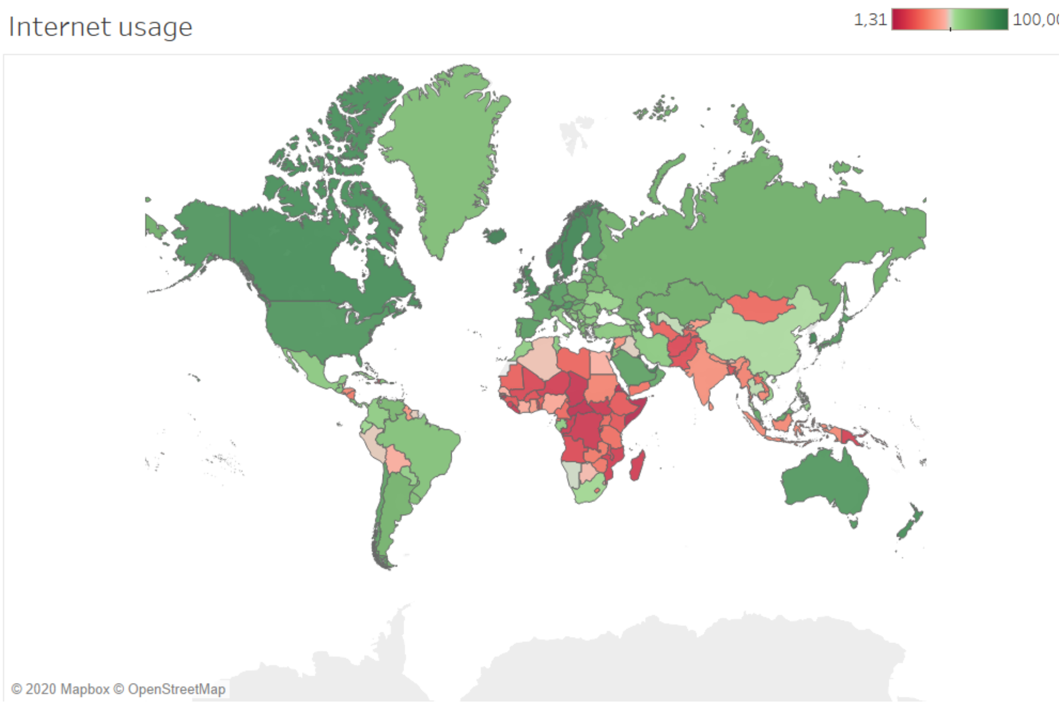


Figure 3.27: Distribution of the internet usage

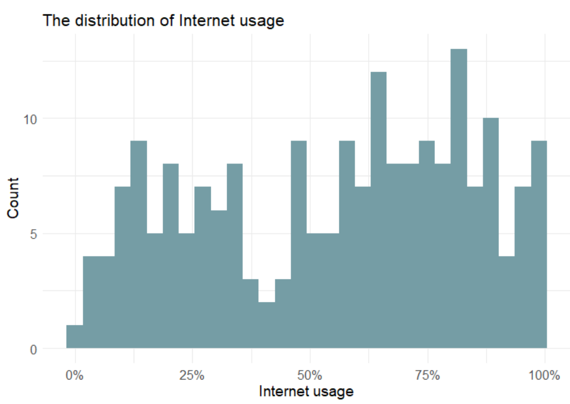


Figure 3.28 indicates the distribution of the cybersecurity index across countries. Similarly to previous maps, the developed world is linked to higher index values while African countries and south Asia struggle with security. Figure 3.29 reveals the distribution of the variable. There was a cluster of countries with poor security that correspond to the African countries.

Figure 3.28: Distribution of the cybersecurity level across countries

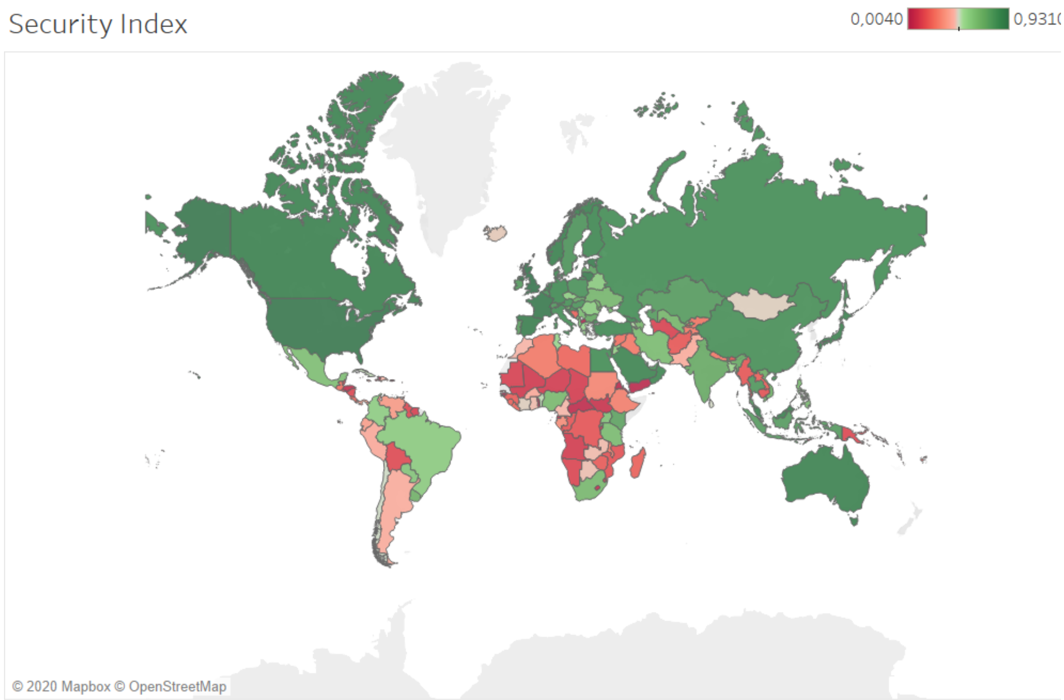
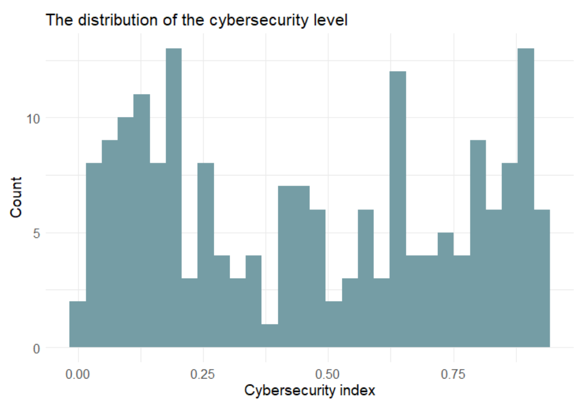


Figure 3.29: Distribution of the cybersecurity index



As for the dependent variable, Figure 3.30 reveals the distribution of the similarity of a username and a password. People in developed countries (e.g., North America, Europe) use a dissimilar username and a password, while people in poor and developing countries (e.g. African states, parts of Asian states) use passwords similar to their usernames. Surprisingly, it seems to be less related to the spoken language than the length (for the password length, see next page).

Figure 3.30: Distribution of the similarity of a username and a password across countries

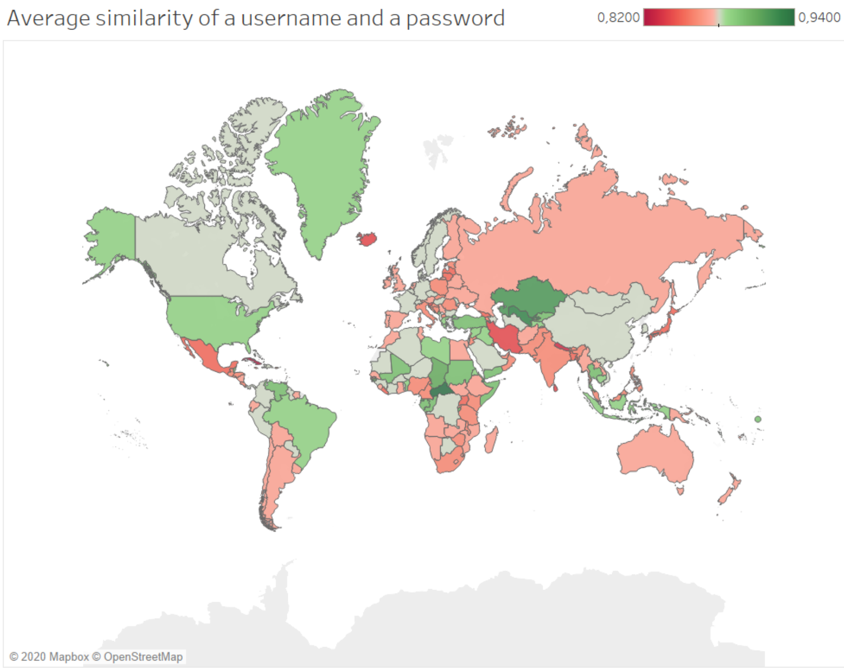


Figure 3.31: Distribution of the similarity of a username and a password

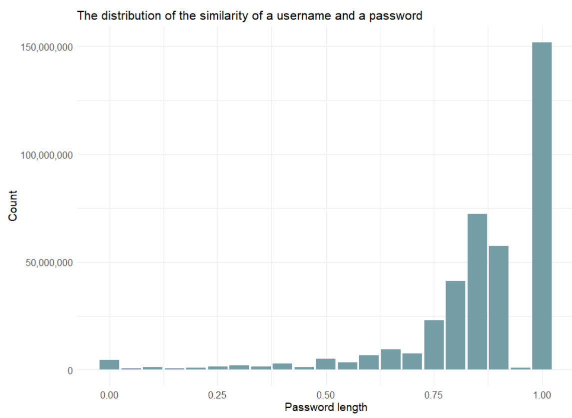


Figure 3.32 reveals the distribution of a password's length across countries. The map suggests a correlation with the language spoken. For example, South American countries where Spanish is dominating have, on average, longer passwords than the Brazilian population, where Portuguese is the most common language. Furthermore, German is known for concatenating words, resulting in longer words than, for example, in English. It can be seen that in Europe, German passwords seems to be longer than passwords in surrounding countries.

Figure 3.32: Distribution of the length of a password across countries

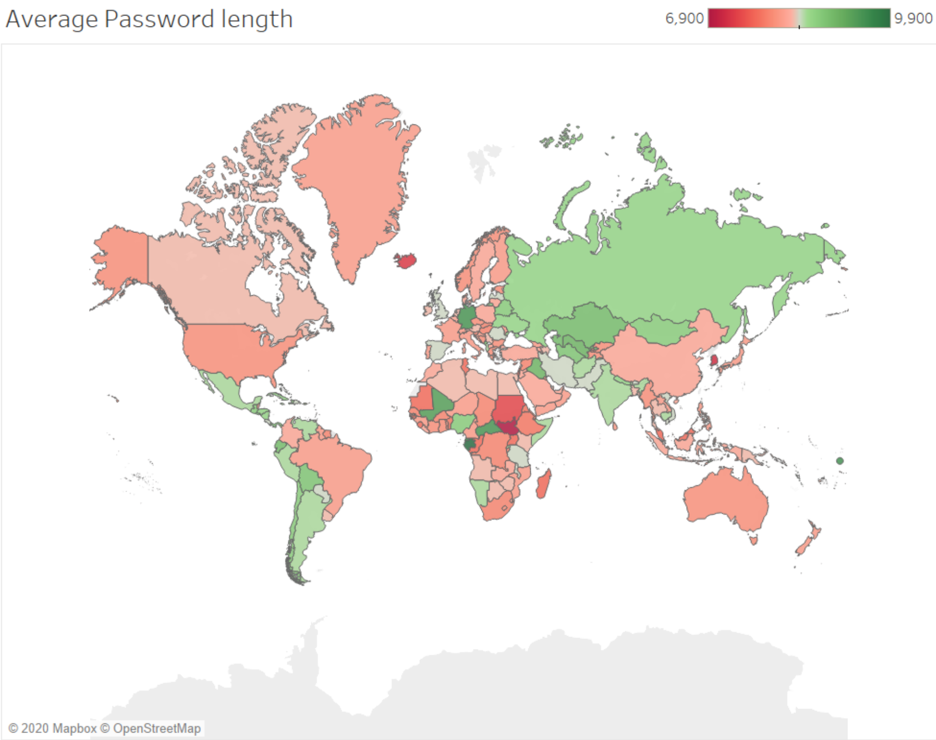
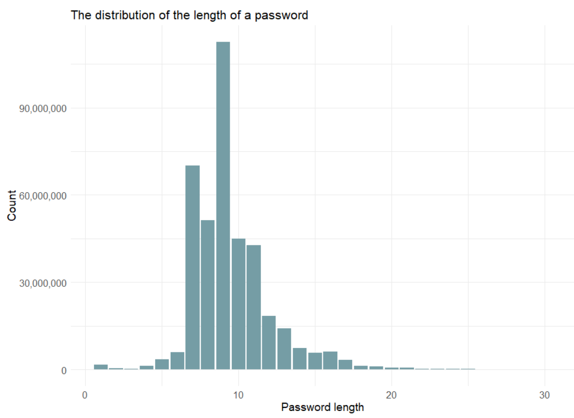


Figure 3.33: Distribution of the length of a password



Tables A.8, A.7, A.6, A.5 and A.4 in the Appendix show a detail descriptive statistics for Macroeconomic variables, the Effort, the Password length, the Similarity of a username and a password and the gender respectively.

Descriptive statistics of Model 2 data

This dataset consisted of users that appeared multiple times. Furthermore, if a user appeared more than twice, a random pair was chosen.

Macroeconomic variables on the country level were identical to Model 1 data and can be inspected in the previous section. Additionally, the password length and the Effort are not relevant in this model, and thus, they were omitted.

On the contrary, gender was still relevant. As the dataset's construction differed from the first one, Table 3.21 reveals the percentage of males and females in the sample. While the number of observations was significantly smaller than for Model 1, the distribution of males and females remained similar.

Table 3.21: Distribtion of the gender in the dataset for Model 2

Males	Females	Unknown	Count
16.6%	12.9%	70.5%	55 198 466

In this model, the variable to explain was the similarity between two passwords of the same user. From the methodological point of view, it was calculated using the same Levenshtein Distance as in Model 1. Table 3.22 reveals the sample statistics of the target for Model 2. The statistics suggest that there were both equal and completely distinct password. The mean suggests that, on average, some changes are required to derive one password from another.

Table 3.22: Sample statistics of the Password similarity

Min	Mean	Max	SD
0.00	6.61	30.00	4.01

Figure 3.35 reveals the distribution of the password similarity across countries. Surprisingly, users in African countries seem to use different passwords. However, there is a decent variability in the figure. Users in Russia seem to use more similar passwords than users in European countries and America. Overall, their similarity of passwords seems to be stronger than the similarity between a username and a password.

Figure 3.34: Distribution of the similarity of passwords across countries

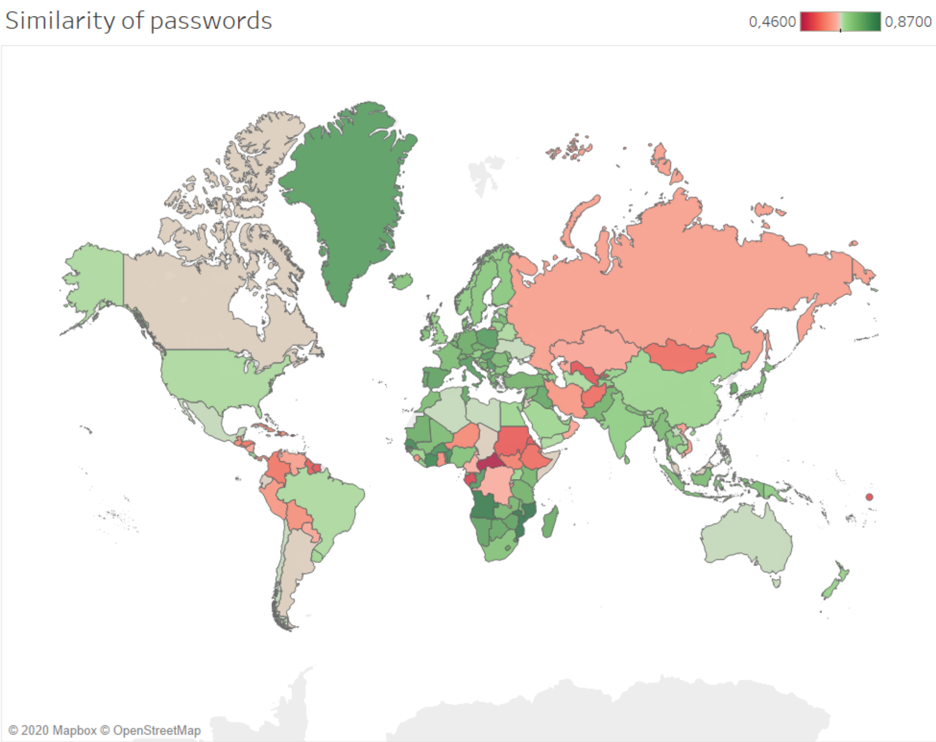
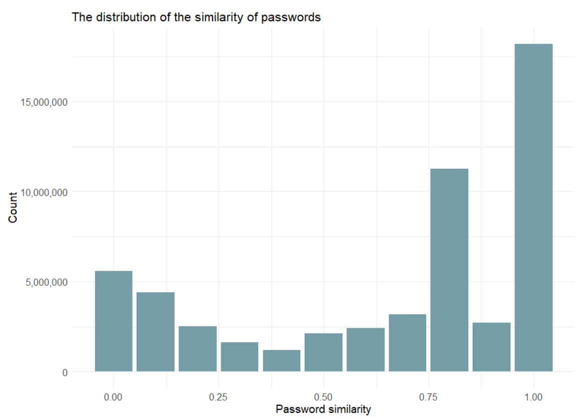


Figure 3.35: Distribution of the similarity of passwords



3.5.2 Description of the sampling strategy

The resulting sample's size was a success on the one hand and an obstacle on the other. Hundreds of millions of observations were too large for a decent running time of models. The text file had more than 20 gigabytes, which did not fit into the memory, and even batch processing would take a significant amount of time.

Thus, it was decided to perform sampling to decrease the data's size to an acceptable size while preserving the information necessary for models. Given the size of the data set, even an aggressive sampling would imply more than enough observations.

Model 1 sampling

In the first model, it was aimed to describe the similarity between a username and a password. In the cleaned sample, one user might appear multiple times. Consequently, it was decided to randomly select one observation per user to avoid any user implied bias (some users appeared more than a thousand times).

This resulting dataset was still too large, and thus, it was decided to perform sampling. As discussed previously, it was suspected that some countries might systematically behave differently from others. Given the number of observation, a random sample that would contain 1% of all observations could omit some TLDs as a few of them contain only dozens of thousands of observations. To ensure that the modelling sample contains all TLDs, it was decided to perform a stratified sampling.

Stratified sampling is a sampling method where the population is divided into subpopulation for further sampling. To ensure that the distribution of TLDs in the sample of the original data is under control (i.e. preferably all TLDs should be included), stratified sampling was performed through TLDs. Note that if the hypothesised model would not include TLD fixed effects, random sampling could be performed.

During the stratified sampling, the observations were sampled in proportions related to the countries' population. That means the sampling follows the real distribution of users among countries. If, for example, Germany has around 83 million inhabitants and, for example, the Czech Republic has around 10 million inhabitants, it would be convenient to have around 8 German users per 1 Czech user.

In addition to that, if no country fixed effect would be expected, this sam-

Table 3.23: Demonstration of the stratified sampling

Country	Population	Scaling ratio	# of observations
The Czech Republic	10 669 709	1.00	10 000
Germany	83 132 799	7.79	77 915
Spain	47 076 781	4.41	44 122
Slovaquia	5 454 073	0.51	5 112
Afganistan	38 041 754	3.57	35 654

pling strategy should not cause harm to the data distribution as it remains random, and by implication, the selection based on the country should not distort the sampling.

The population data were obtained from the 2019 Revision of World Population Prospects prepared by the United Nations [19]. The Czech Republic was chosen as the base country. For the rest of the countries, a scaling coefficient was computed to calculate the number of observations per country while retaining the populations' ratios. Table 3.23 demonstrates this stratification on a few countries.

To ensure that the sampling produces consistent results; first, the sample was derived using two different random seeds and second, two base sizes of the sample were taken into consideration—10 000 Czech users and 20 000 Czech users.

Following figures demonstrate the comparison of sample statistics among identified strategies.

Model 1 - Password length Figure 3.36 reveals the distribution of the Password length among the three samples. As one can see, the sample statistics are almost identical regardless of the statistical measure (i.e. minimum, mean, median, maximum and standard deviation). These findings suggest that the sampling should not significantly harm the data.

Model 1 - The Effort Figure 3.37 describes how the sample statistics for the Effort varied under the different sampling strategies. One can see similar statistics for all three samples. A minor deviation exhibits the sample with random seed 1 000 with 100 thousand observations. Nevertheless, the difference is less than one percentage point in all four categories.

[19] Available at https://population.un.org/wpp/

Figure 3.36: Sample statistics of the Password Length

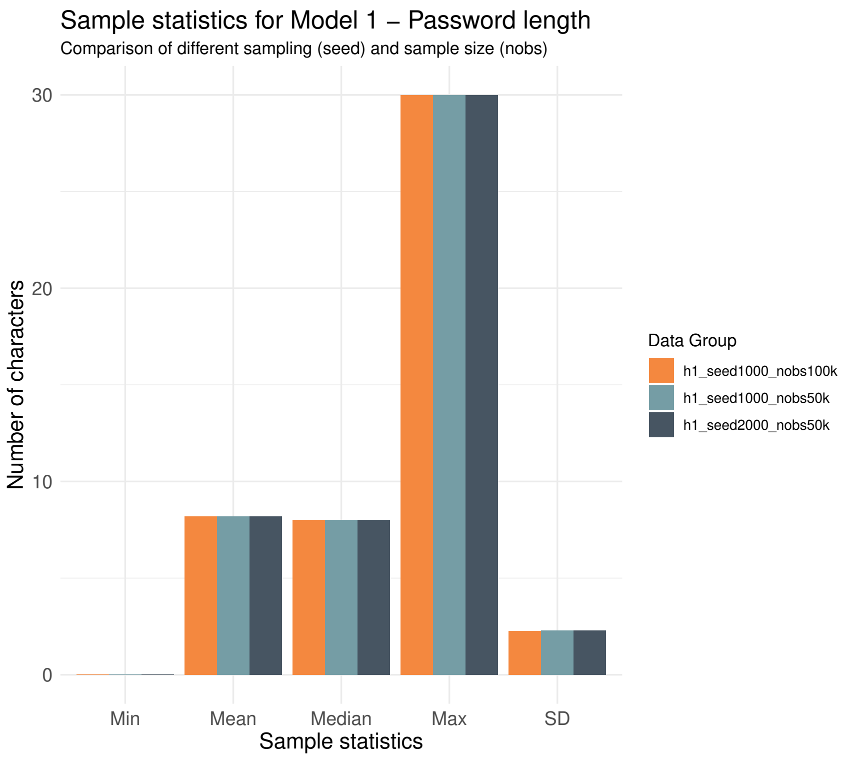
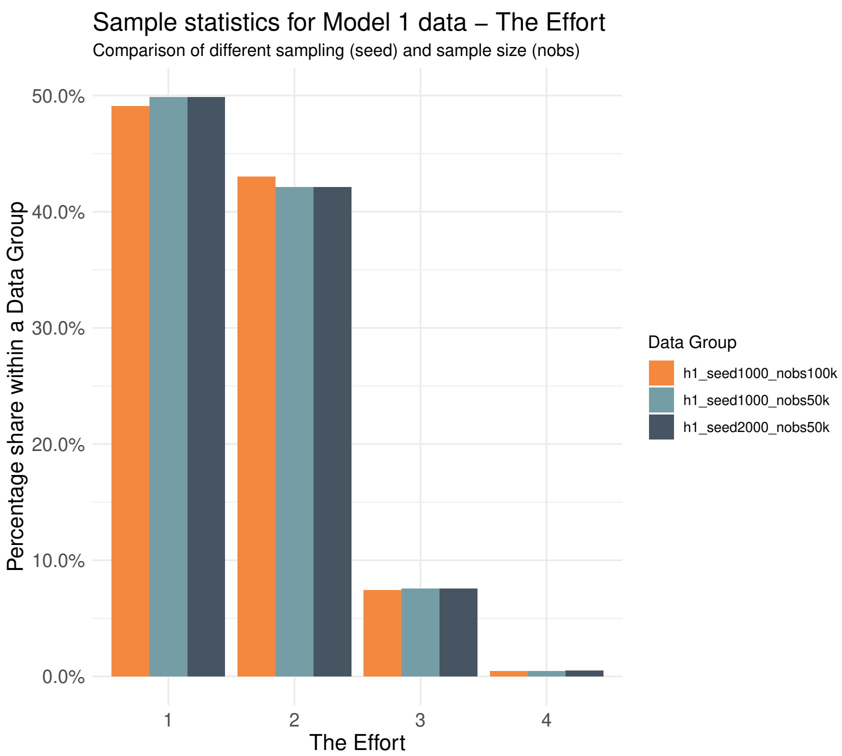
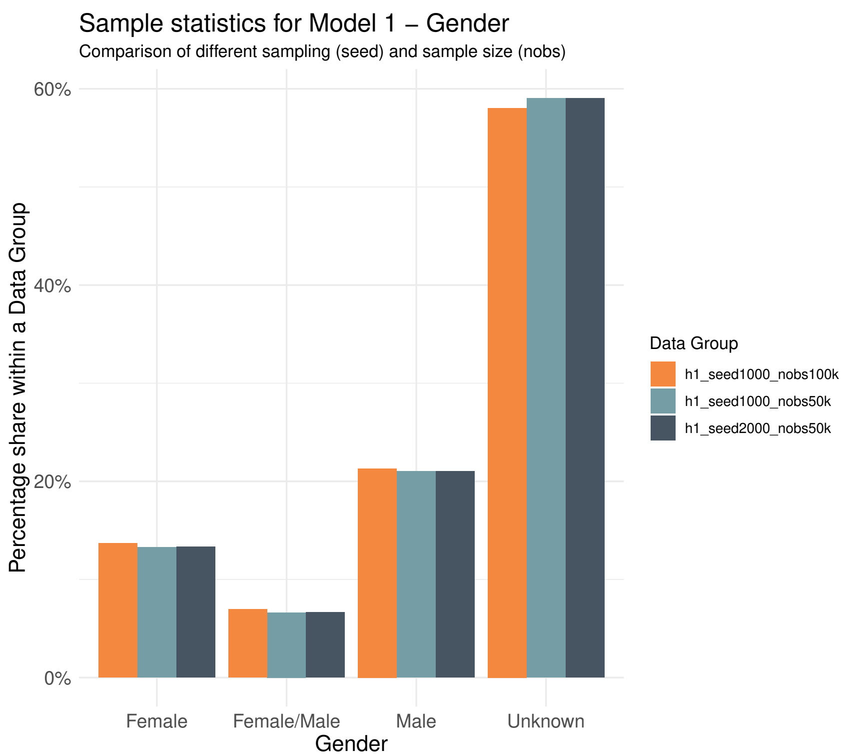


Figure 3.37: Sample statistics of the Effort



Model 1 - Sex Speaking about the sex, Figure 3.38 indicates the differences in the gender distribution among the sampling strategies. All three samples contain a similar share of identified gender. The differences are well below one percentage point, and thus, it has been concluded that the sampling strategy seems to be reasonable.

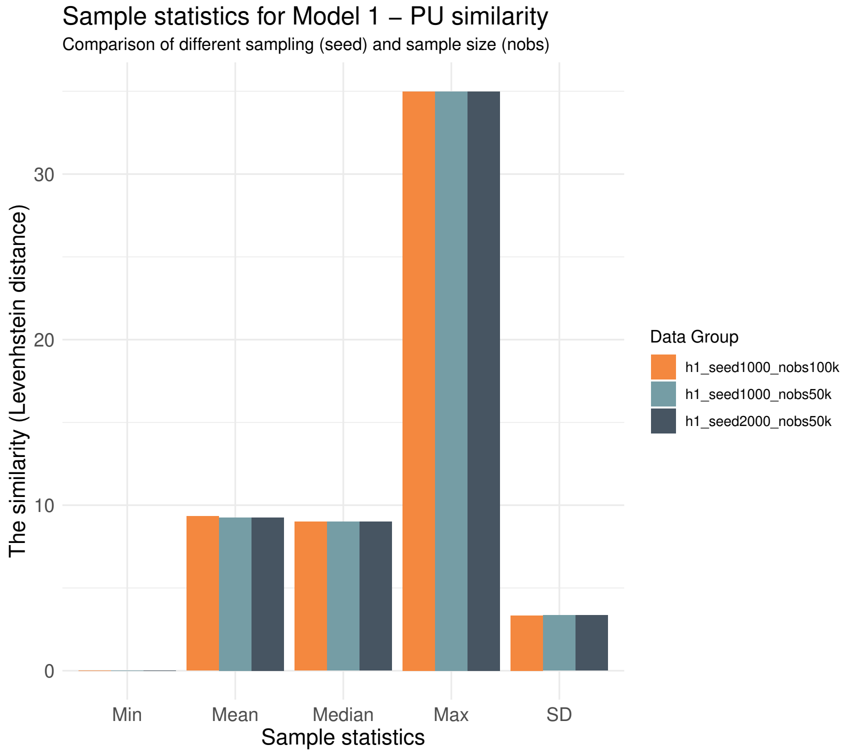
Figure 3.38: Sample statistics of the Sex



Model 1 - Password Username Similarity Figure 3.39 reveals how the target variable, the similarity between a username and a password, differ among the samples. It can be observed that there are negligible differences in the sample statistics among all three sampling strategies. That is a positive finding, and it might be concluded that the sampling strategy should not significantly affect the analysis.

The distribution of macroeconomic variables did not change with the sample size and seed because the country ratios were maintained, and regardless of what observations one would choose, they would be identical within a country.

Figure 3.39: Sample statistics of the PU Similarity



3.5.3 Analysis of the relationship of the independent variables

This subsection is dedicated mainly to the analysis of interactions among the variables. The analysis might suggest potential multicollinearity issues.

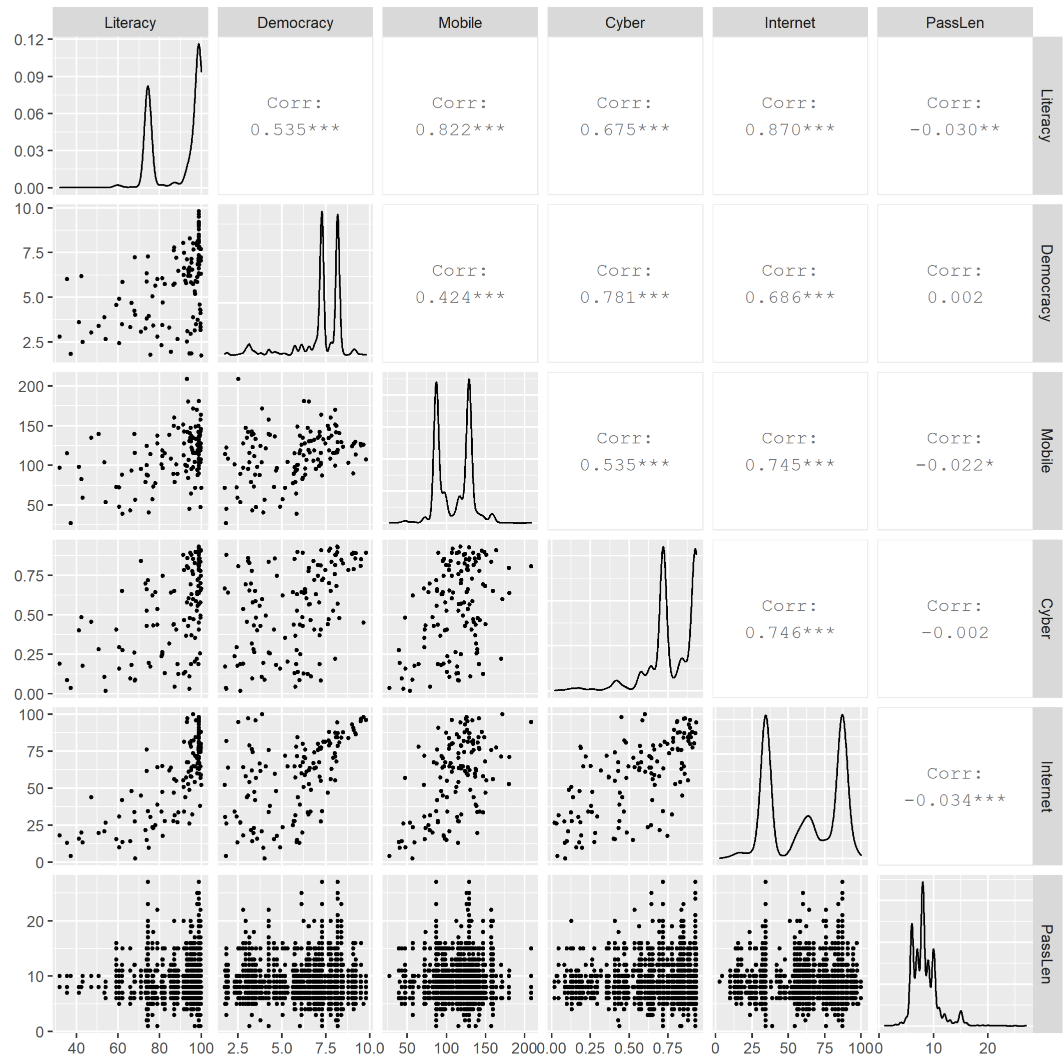
Correlation analysis - Model 1 data

Figure 3.40 reveals the relationship among the numerical variables. Macroeconomic variables and the Password length were relevant to the analysis. As one can see, the results suggest a couple of highly correlated pairs of variables.

The correlation analysis reveals three highly correlated variables with a correlation coefficient higher than 0.7. For a more straightforward analysis of these pair of variables, Figure 3.41 is a ranked list of the correlation coefficients of all pairs within this analysis.

As one can see, the two highest correlation coefficients are related to the Literacy variable. Furthermore, the Internet variable is related to the third and fourth highest variable. These pairs of variables might cause multicollinearity

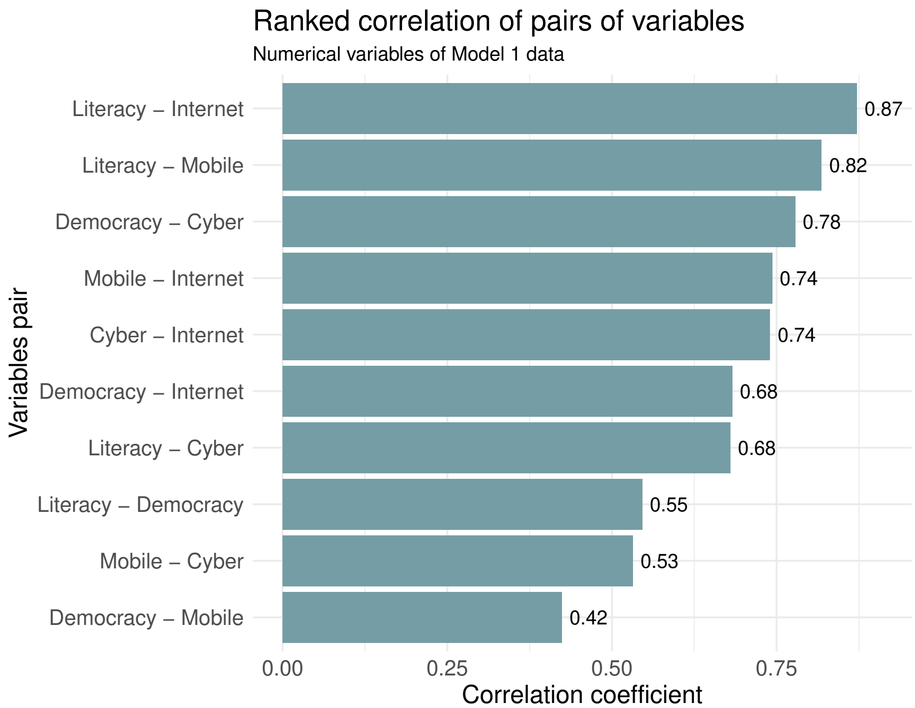
Figure 3.40: Correlation analysis of continuous variables



issues, and thus, these findings have to be reflected in the regressions. The rest of the correlation coefficients is lower than 0.51.

As described in the sampling strategy, three samples for Model 1 was generated. The correlation analysis was done on one of the samples (seed 1000 with base scaling 10000). Figure 3.42 shows a comparison of the correlation coefficients of pairs of variables among the three samples.

Figure 3.41: Ranked correlation of Model 1 variables



As one can see, the correlation coefficients do not differ dramatically across the samples. That is in line with the previous analysis of the sampling strategy. The samples seem to have similar properties.

Correlation analysis - Model 2 data

Figure 3.43 reveals the relationship among the numerical variables for Model 2 main sample. Only macroeconomic variables were relevant to the analysis. As one can see, the results suggest a couple of highly correlated pairs of variables.

The correlation analysis reveals one strongly correlated pair of variables - Internet and Literacy. The rest of the pairs seems to be uncorrelated or only modestly correlated. The correlation coefficient goes up to 0.63, which might cause an issue in the model estimation.

Figure 3.42: Comparison of correlation coefficients on three samples

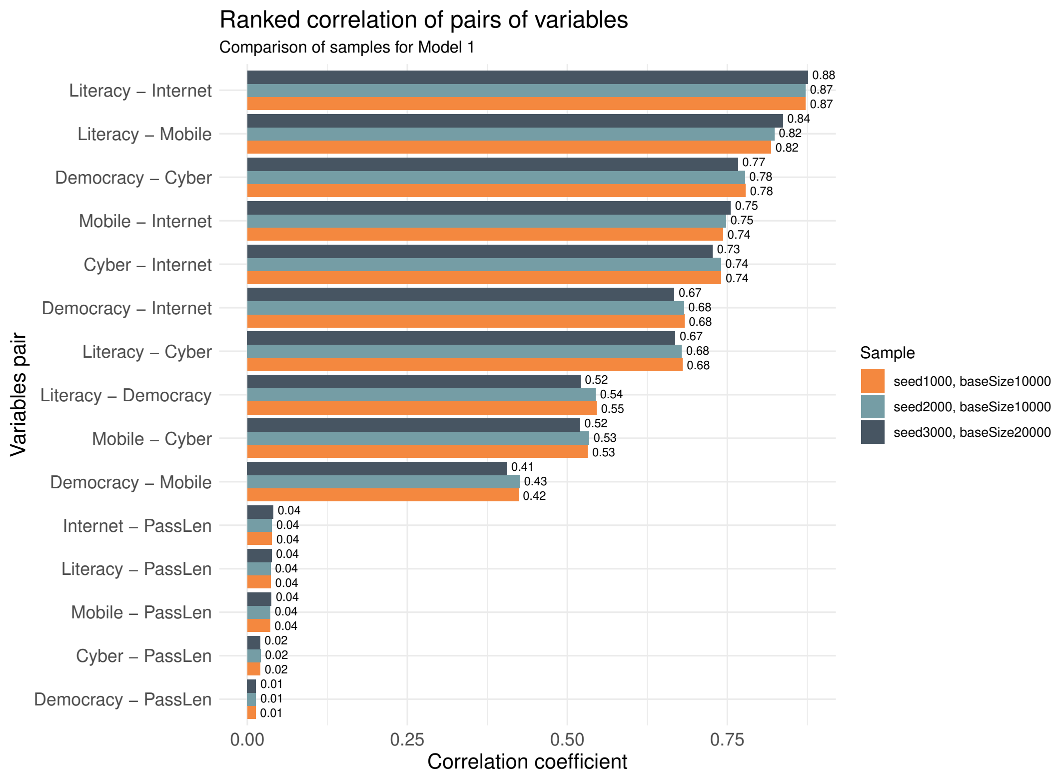


Figure 3.44 shows the ranked correlation coefficients of pairs of variables. The most correlated pair is Literacy - Internet, with a correlation coefficient of 0.86. The second highest pair is Literacy - Mobile, with a correlation coefficient of 0.63. That is, the Literacy variable seems to be causing most of the high correlation. Mobile - Internet with a correlation coefficient of 0.56 is the third one. Model 2 should be modified accounting for these potential sources of multicollinearity.

Last, Figure 3.45 reveals a comparison of the correlation of pairs of variables among the three samples. It can be observed that all three samples have similar correlation coefficients of the pairs of variables. However, minor differences appear in the case of Cyber - Mobile and Mobile - Internet pairs. That is not unexpected, given the sampling strategy. The important conclusion is that the correlation coefficients seem to be identical for seed 1 000 and 2 000 with a base of 10 000.

In conclusion, the correlation analysis suggest pairs of highly correlated variables for both Model 1 and Model 2 that might be the source of multicollinearity. The modelling step should account for that.

Figure 3.43: Correlation analysis of continuous variables

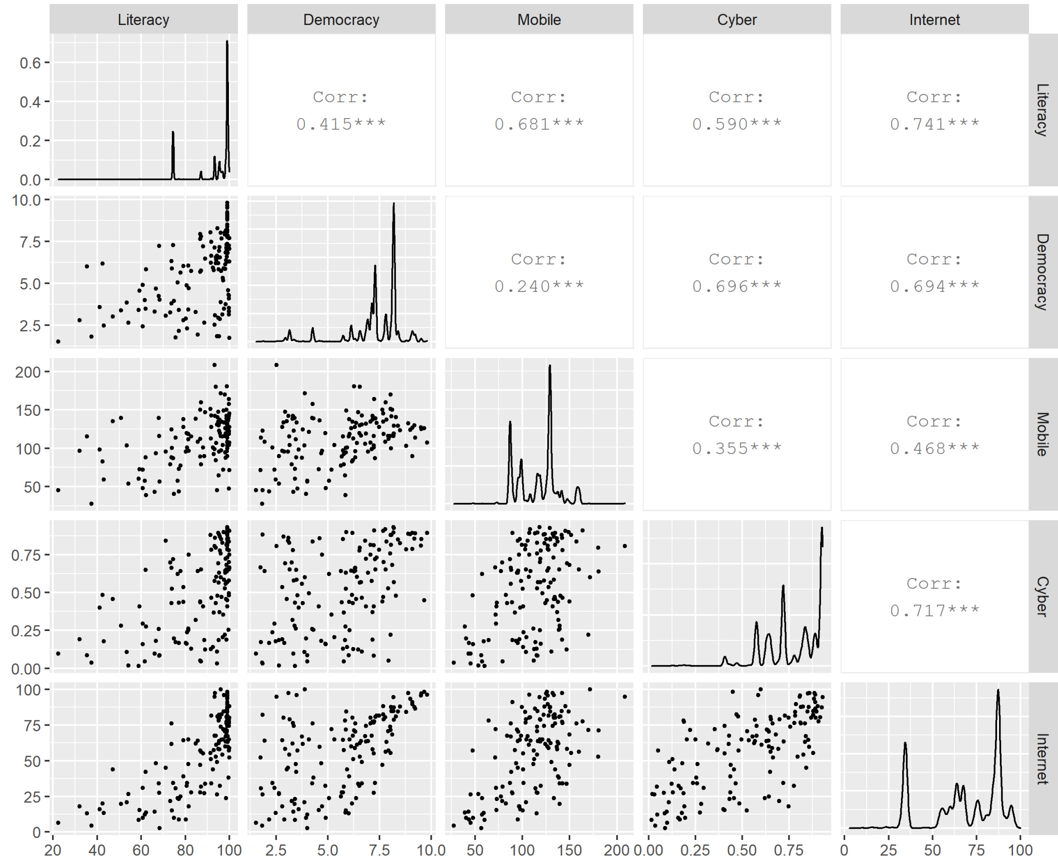


Figure 3.44: Ranked correlation of Model 2 variables

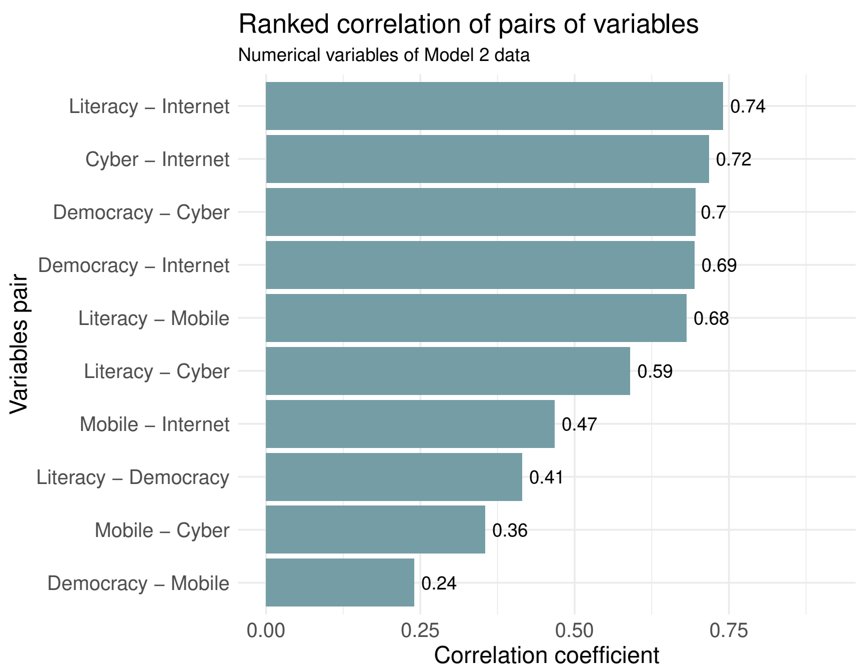
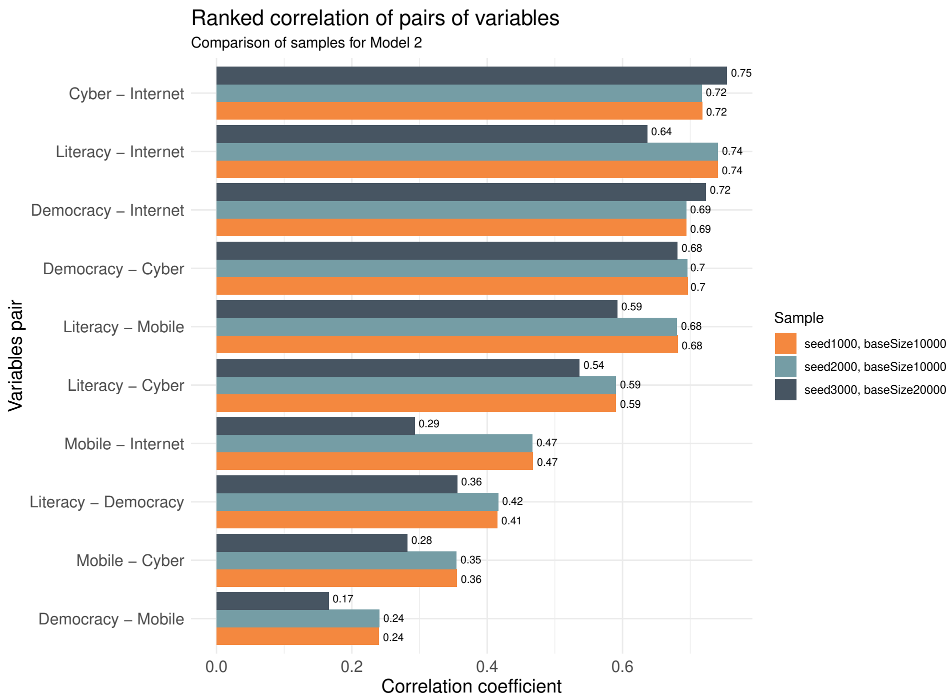


Figure 3.45: Comparison of correlation coefficients on three samples



3.5.4 Analysis of the variable's relationship with the target

This section should outline an overview of the predictors' relationship with the target variable for both Model 1 and Model 2.

PUSim and Predictors - Model 1

Figure 3.46 informs about the relationship between the Password-Username Similarity and the predictors relevant for Model 1. In regards to numerical variables, Password length suggests a positive relationship with the Password-Username Similarity. On the contrary, charts with the macroeconomic variables do not suggest a strong relationship. There seems to be much noise in the data.

The relationship with macroeconomic variables is not surprising as they capture the nation's behaviour, and the dataset is held on the user level. Nevertheless, the regression analysis might reveal some patterns as there are many observations (and they might overlap on the chart).

Regarding the categorical variables, according to the boxplot, the Sex variable seems to not play a significant role in the password derivation. As expected, the chart with the Effort suggests a positive relationship between the PUSim and the Effort. Users, including multiple character types in their password, tend to have lower similarity between their password and username.

The Sentiment variable also suggest a pattern. The chart suggests that users whose passwords could be assessed as having positive connotations have slightly higher PUSim. That is, their passwords are less similar. In any case, the difference seems to be very small.

Additionally, users having a password with neutral connotations seems to have higher PUSim, that is, having passwords less similar to their usernames. That might suggest that users that have passwords with positive (or negative) connotations might have part of this connotation in the username as well. Nevertheless, it might also mean that they might use a part of their username without having a username with any of the polarity for users with some polarity in their passwords.

PPSim and Predictors - Model 2

Figure 3.47 shows the relationship between the Password-Password Similarity and the predictors for Model 2. The numerical variables exhibit similar patterns to the Model 1 data inspection. A large number of (overlapping) points make it difficult to make a conclusion using the visual inspection.

As for the categorical variables, only Sex is used for Model 2. Similarly to the Model 1 inspection, the chart does not suggest a strong difference between Males and Females.

Figure 3.46: Relationship between PUSim and the variables

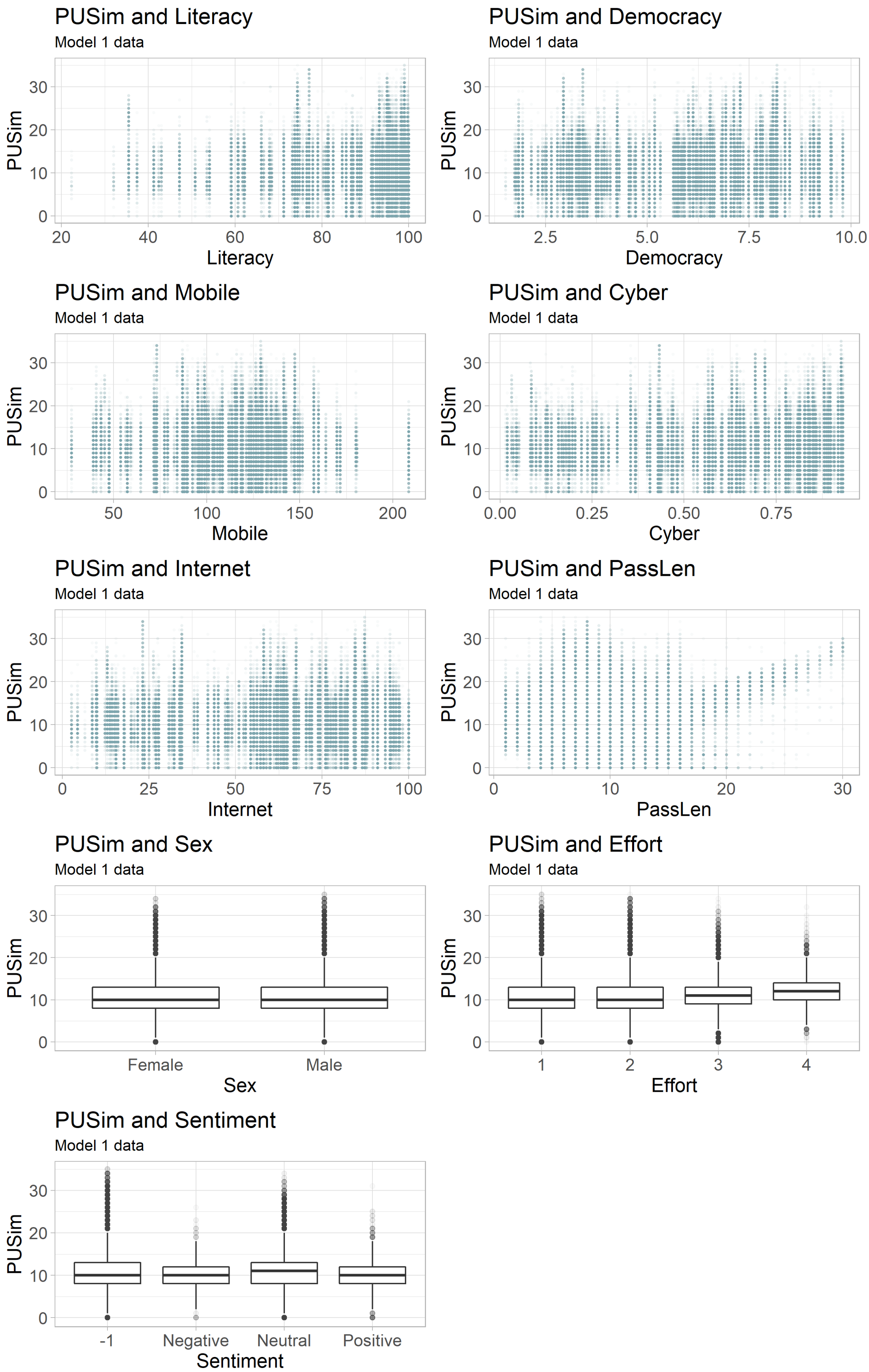
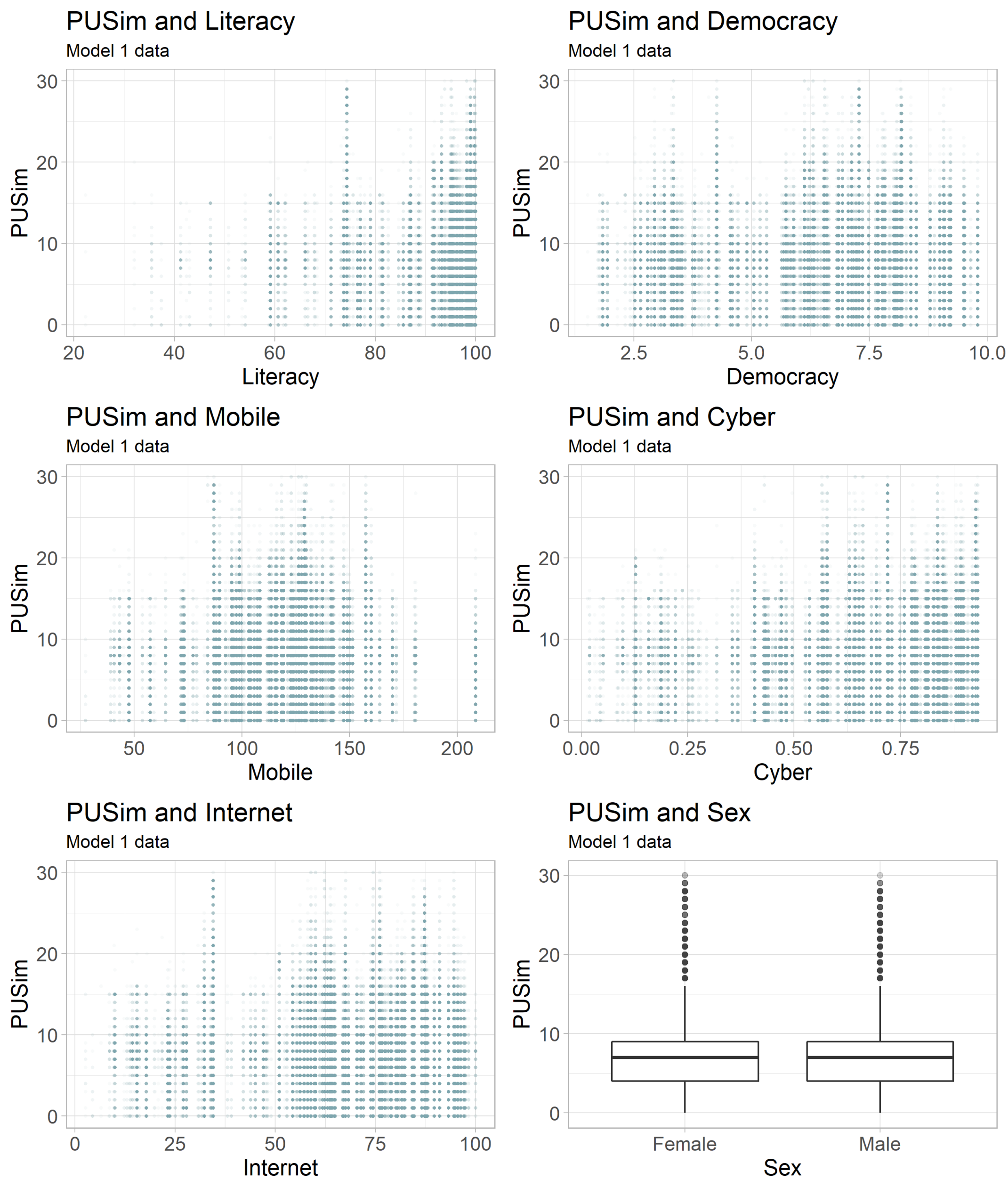


Figure 3.47: Relationship between PPSim and the variables



Chapter 4

Results

In this chapter, one can find an overview of the estimated relationships. Nine models corresponded to the hypothesized model 1, and 5 models corresponded to the Model 2 family. Models were estimated using the VGAM package in R[1].

The generalized ordered logistic regression does not rely on many assumptions. The outcome was expected to be ordered, the sample size was large, predictors were either continuous or binary with non-zero variance, and the potential multicollinearity was mitigated in the previous section.

The generalized ordered logistic regression might generate an uncomfortably large number of estimates. For example, having 30 levels of the target variable and five predictors would lead to more than 150 estimates.

That makes the analysis more complex. First, the number of estimates to be evaluated is large. Second, one has to pay attention to the evolution of β coefficients through the different cutoffs.

As a consequence, the models are presented using the goodness of fit summary and charts informing about the β_i^j coefficients, including the significance and confidence intervals. A more comprehensive but more extended summary can be found in the appendix.

4.1 Model family 1 - the similarity of a username and password

This model family aimed to explain why users use similar usernames and passwords as discussed before.

[1] https://cran.r-project.org/web/packages/VGAM/index.html

Table 4.1: Summary of Model 1 family

	Model (m1_)							
Variable name	full	base	seed	size	PCA	sent	TLD	lan
PassLen	x	x	x	x	x	x	x	x
Cyber	x	x	x	x	x	x	x	x
Mobile	x							
Effort2	x	x	x	x	x	x	x	x
Effort3	x	x	x	x	x	x	x	x
Effort4	x	x	x	x	x	x	x	x
SexF	x							
PCAs					x			
Polarity						x		
TLDs							x	
Languages								x

Table 4.2: Goodness of fit measures for hm1_full model

Model	*Deviance*	*LogLikelihood*	*Degrees of freedom*
hm1_base	999 684.8	-499 842.4	2 117 783

Table 4.1 gives an overview of the estimated models. One can see the connection between a model name suffix (e.g., full or base) and variables included. For example, *m1_full* indicates Model family one and the full version of the model.

4.1.1 Model family 1 - initial model (m1_full)

The first model gives a general overview of the estimated β coefficients of the chosen variables. It includes all the variables with expected impact, excluding variables implying potential multicollinearity. The Password-Username similarity is being explained by the Password length, Cybersecurity index, Mobile usage, sex, and The Effort made.

Table 4.2 informs about the goodness of fit of the model. The deviance is 999 685, loglikelihood is -499 842 and degrees of freedom are large, more than 2 100 000.

Figure 4.1 gives an overview of the estimated β coefficients. In the charts, each observation corresponds to one of the $P(Y \geq j)$ model. $P(Y \geq j)$ is a different cutoff j, which indicates where the dichotomization of the target categories was performed. The error bars indicate the 95% confidence interval. The asterisks indicate whether the particular estimated β was statistically different from 0 using the following thresholds of p-values: 0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 1'. This will hold for the rest of the model related charts as well.

PassLen The results suggest that the Password length significantly affect the Password-Username similarity. The estimates are statistically significant for most of the cutoffs j. However, for larger cutoffs j, the estimates cease to be statistically different from 0 (from cutoff 25).

The magnitude and direction of the estimated coefficient evolve across the cutoff j. That makes the interpretation slightly more complicated. On two sets, cutoff 1 to 5 and cutoff 15 to 29, the estimated coefficients are negative, indicating a negative effect on the probability of being at or above a cutoff j. Between cutoff 6 and 14, the estimated coefficients are positive, indicating a negative effect on being at or above a cutoff j. The confidence intervals suggest the coefficients are mostly statistically different (at the 5% significance level) except the high cutoffs, where the coefficients are not significant.

The magnitude of the coefficients varies from less than -0.10 to nearly 0.10. An additional character in a password increases, on average, the log odds by up to approximately 0.1, ceteris paribus.

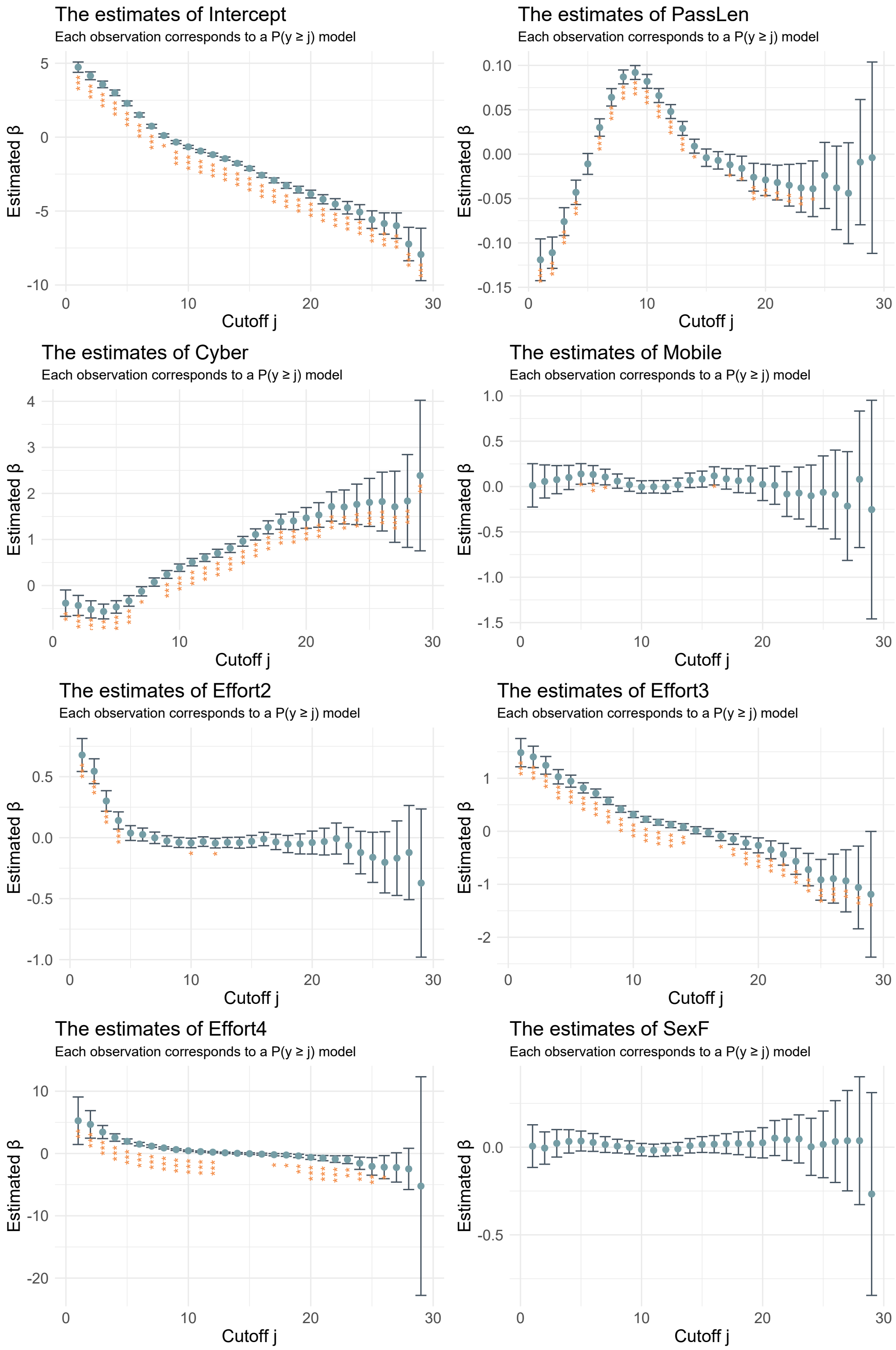
These results suggest that, on average, for highly similar Passwords and Usernames (low cutoffs j), the Password Length decreases the probability of being at or above a particular level, ceteris paribus. In other words, it decreases the probability of having a less similar password and username. Moreover, that happens as well for highly dissimilar passwords and usernames (high cutoffs). For very different passwords and usernames, the password length decreases the probability of having even more dissimilar password and username.

For moderately similar passwords and usernames (i.e., cutoff 6 to 14), the password length increases the probability of having a more dissimilar password and username pair. For moderately similar passwords and usernames, longer passwords lead to less similar passwords and usernames.

Cyber The results suggest that the Cybersecurity index significantly contributes to the Password-Username determination. An increase in the Cybersecurity seems to be related with an increase in the dissimilarity of a password and a username.

The estimated coefficients are negative for minimal cutoffs j (i.e., cutoff 1 to 7) and positive for cutoff eight and above. The magnitude varies from approximately -0.5 to nearly 2.5. That indicates that on average, a one-unit

Figure 4.1: The estimated β coefficients of m1_full model



*Asterisk in a chart indicates whether the estimated β coefficient is statistically different from 0 using following thresholds of the p-values: 0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 1'* increase in the Cybersecurity index would lead to up to nearly 2.5 increase in the log odds of being at or above a cutoff j, ceteris paribus.

In plain words, the findings suggest that for very similar passwords and usernames, the Cybersecurity index is related to a higher similarity between a username and a password. However, for moderately to highly dissimilar passwords and usernames, the index contributes positively to their dissimilarity, indicating better security practice.

The findings are in line with what was expected. The Cybersecurity environment seems to be associated with password management. Better Cybersecurity level seems to lead to the higher dissimilarity between passwords and usernames.

Mobile The results suggest that the mobile usage is not significant for explaining the Password-Username similarity.

The estimated β coefficients oscillate around 0 with relatively large 95% confidence intervals. Only four of the estimates are significantly different from 0 at 5% significance level.

In the summary, it was failed to find evidence that Mobile usage would be a significant determinant of the Password-Username similarity.

The Effort The Effort made by the users was composed of three dummy variables, Effort2, Effort3 and Effort4. The estimated β coefficients of The Effort2 seems to be partially helpful for explaining the Password-Username similarity. On the contrary, the results indicate that both The Effort3 and The Effort4 contribute significantly to the Password-Username similarity.

The first four β estimates of The Effort2 are significant. The rest of the coefficients is not significantly different from zero. Estimates oscillates around 0 with large 95% confidence intervals (except two cutoffs).

The first four β estimates are positive, significantly different from zero with a decreasing magnitude. The first one is more than 0.5 while the fourth one is less than 0.25. The log odds would increase by only 0.5 compared to the base group - using one character only.

The results suggest that for very similar passwords and usernames, using two different character sets in the passwords is related to a higher probability of having less similar passwords and usernames than using only one character set. On the contrary, there was no statistically significant evidence to support this claim for medium similar or dissimilar passwords and usernames.

The estimated effect of The Effort3 seems to be significant across the cutoff j, and nearly all of the estimates are statistically different from 0. There is a decreasing trend in the estimates. They range from 1.5 (for the most similar passwords and usernames) to more than -1 (for the least similar pairs). That means that the log odds of being in a higher group for the most similar passwords and username increases, on average, by 1.5, ceteris paribus, in comparison with the base group consisting of only one character set.

These findings indicate that for highly similar passwords and usernames, using three different groups of characters in a password is related with a lower similarity between the password and the username. On the other hand, for highly dissimilar passwords and usernames, the three-character sets in the passwords are associated with higher similarity between a username and a password.

The estimated β coefficients of the Effort4 and the Effort3 are similar. They are mostly statistically significant with a decreasing trend in magnitude. The estimates range from more than 5 to -5. Nearly half of the estimates are statistically different from 0.

As in the case of The Effort3, for highly similar passwords and usernames, the structural diversity of the password is associated with less similar passwords and usernames, compared with only one character set in the password. On the other hand, for dissimilar passwords and usernames, the diversity of the password is associated with more similar passwords and usernames. Nevertheless, only a few of the coefficients in this region are statistically significant.

The decreasing trend of The Effort2 and The Effort3 might be explained from two perspectives. First, for very similar passwords and usernames, if a user does not care about the variety of characters used, he might not care about deriving the password from the username (or vice-versa). Second, for dissimilar passwords and usernames, if a user prefers a diverse password, it would be more likely to have a slightly more similar password and username.

SexF The results suggest that the sex does not have a significant effect on the Password-Username similarity.

All the estimated β coefficients corresponding to the gender are statistically insignificant, being indifferent to zero at 5% significance level.

In conclusion, no evidence was found to support the hypothesized effect of gender on the Password-Username similarity.

To sum it up, Cyber, the Effort, and Password length seem to play an impor-

Table 4.3: Goodness of fit measures for hm1_base model

Model	*Deviance*	*LogLikelihood*	*Degrees of freedom*
hm1_base	685 245.7	-342 622.8	1 310 684

tant role when explaining the password-username similarity. On the contrary, the effect of Mobile and Sex was not confirmed by the results.

Next model reveals how the coefficients of PassLen, Cyber and The Effort will change when Mobile and Sex are excluded from the equation.

4.1.2 Model family 1 - base model (m1_base)

This model includes variables from the initial model where evidence of a significant relationship with the Password-Username similarity was found. That is the Cybersecurity index, password length and the Effort. This model will be used for assessing the effectiveness of the experimental variables.

Table 4.3 reports the Goodness of Fit measure of the model. Deviance is over 680 000, being lower than in the Full model. Loglikelihood is -342 622, being higher than in the Full model. Degrees of freedom is over 1.3 million.

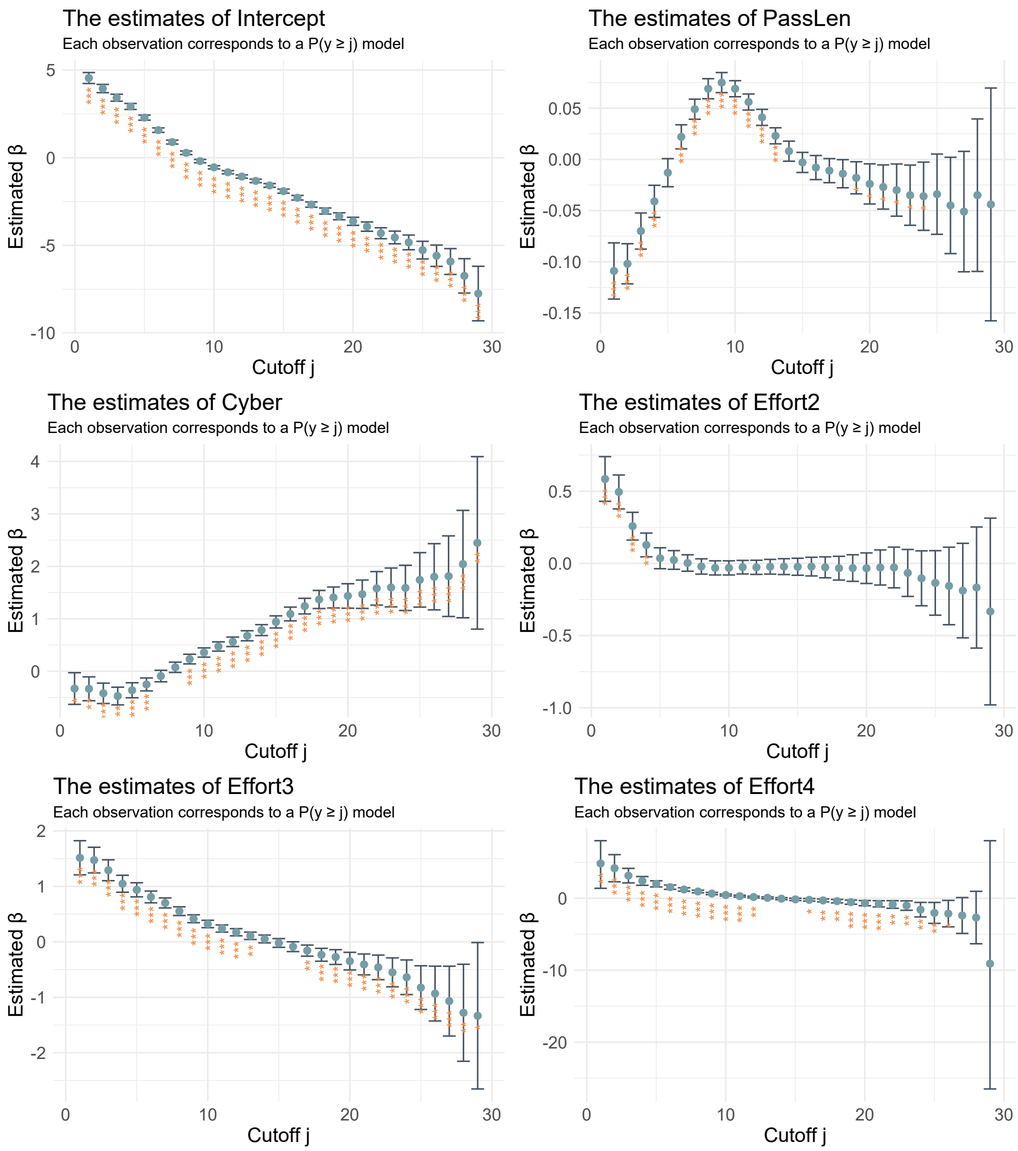
Figure 4.2 gives an overview of the estimated coefficients of the base model. Generally speaking, estimated β coefficients are similar to what was observed in the full model earlier. Minor differences are observed, which is not a surprise, as the variables in the model do not have perfect explanatory power and the parameter space for the optimization was very complex.

PassLen The password length remain primarily significant (mainly for the first half of the cutoffs j), and the estimated coefficients exhibit a similar shape to what has been seen in the Full model. No dramatic change occurred after the exclusion of redundant variables.

Cyber Cybersecurity remains strongly significant with nearly monotonous estimates through the cutoff j. The range of the coefficient is nearly identical as well. It is reasonable to believe that the exclusion of variables did not impact significantly the estimated β coefficients of the Cyber variable.

The Effort The Effort2 was retained in the equation because of the complementarity with The Effort3 and The Effort4. AS expected, Effort2 remains insignificant. The Effort3 exhibits a similar pattern as seen before, having a

Figure 4.2: The estimated β coefficients of m1_base model



*Asterisk in a chart indicates whether the estimated β coefficient is statistically different from 0 using following thresholds of the p-values: 0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 1'* significant positive effect for similar passwords and usernames and a relatively strong negative effect on the Password-Username similarity for highly dissimilar ones.

The estimates of Effort4 also suggest similar results to the first model. Diversity of a password leads to higher password and username dissimilarity, mostly for similar passwords and usernames. On the other side of the spectrum, one can describe the highly dissimilar passwords and usernames. Here, the diversity is linked with an increase in the similarity of passwords and usernames.

Cybersecurity and Mobile usage were not the only macroeconomic variables considered for explaining the Password-Username similarity. Unfortunately, due to potential multicollinearity issues, four additional macroeconomic variables were omitted. It would be interesting, though, if these variables contain some information that would explain Password-Username similarity. The next part is devoted to such an investigation.

4.1.3 Model family 1 - PCA model (m1_PCA)

In order to avoid the multicollinearity issues, Internet, Literacy and Democracy variables were transformed using the Principal Component Analysis (PCA). This step should ensure the elimination of the high correlation between the variables. On the other hand, it does not imply that the transformed variables would be useful for determining the Password-Username similarity.

The PCA was done using the built-in *prcomp* command in R. Data were scaled and centred beforehand. Figure 4.3 reveals what proportion of the variance was explained by each principal component. As one can see, nearly 97% of the variance is being explained by the first component. The second component accounts for almost 3% of the variance, and the last component seems to be marginal, accounting only for 0.09% of the variance.

As the purpose of PCA in this analysis is not to decrease the number of variables but rather to bypass the potential multicollinearity. All three components were included in the model. However, it is important to keep in mind that the second and the third components are not explaining much of the variance.

Table 4.4 indicates the goodness of fit of the model. The deviance is over 707 000, which is more than in the case of the base model. Loglikelihood is -353 591, which is slightly lower than in the base model.

The estimated effects of the base variables (i.e., Cybersecurity, Password

Figure 4.3: Percentage of explained variance per the Principal Component

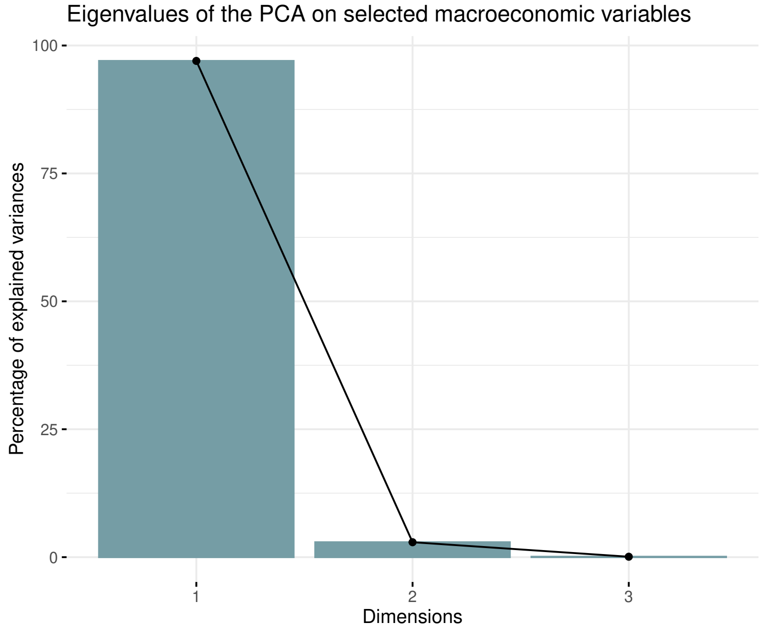


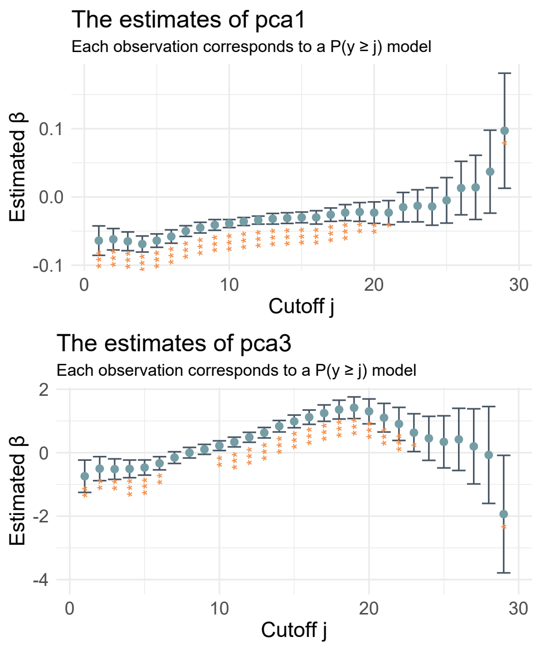
Table 4.4: Goodness of fit measures for hm1_pca model

Model	*Deviance*	*LogLikelihood*	*Degrees of freedom*
hm1_base	685 245.7	-342 622.8	1 310 684
hm1_pca	707 182.5	-353 591.3	1 372 193

length and the Effort) are similar to the base model. Charts and a detailed table with the estimates can be found in the appendix.

Figure 4.4 gives an overview of the estimated β coefficients of the PCA variables. Overall, the PCA related estimates suggest a significant effect on the Password-Username similarity.

Figure 4.4: The estimated β coefficients of m1_PCA model



*Asterisk in a chart indicates whether the estimated β coefficient is statistically different from 0 using following thresholds of the p-values: 0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 1'*

PCA Surprisingly, all three principal components seems to be relatively strongly significant. The effect changes across the cutoff j, indicating asymmetrical effects. This model aims not to report the exact effects of the individual variables but rather to point out that there might be some combined effect worth further investigation.

The PCA1, bearing most of the variance of the Internet access, Literacy rates and Democracy level, seems to have a significant but relatively weak effect on the Password-Username similarity. The first two-thirds of the cutoffs j estimated coefficients are strongly significant at the 5% level. However, their magnitude is small, varies from around -0.6 to around -0.2.

The PCA2 is responsible only for 3% of the variance in the three macroeconomic variables. Around four-fifths of the estimated β coefficients are statisti-

Table 4.5: Goodness of fit measures for hm1_lan model

Model	*Deviance*	*LogLikelihood*	*Degrees of freedom*
hm1_base	685 245.7	-342 622.8	1 310 684
hm1_lan	695 066.8	-347 533.4	1 346 557

cally significant. Compared with estimates of PCA1, the magnitude is slightly larger. Estimates range from approximately -0.2 to 1.2.

Last, the PCA3 related to a negligible amount of variation in the three macro variables, has a significant effect on the Password-Username similarity for half of the estimated β coefficients.

The vital drawback of this model is that the Internet access, Literacy rates and Democracy level together affect the Password-Username similarity significantly and are worth further investigation.

4.1.4 Model family 1 - language model (m1_lan)

One of the goals was to examine whether there structural differences among languages. The former idea was to assign languages to TLDs (countries), make dummy variables, and observe whether there are some statistically significant coefficients. Unfortunately, it was realized it was not feasible.

There were around 50 languages that should have been taken into consideration. An ordered logistic regression with that many predictors and 30 different target levels would imply over 1500 coefficients to be estimated. That was not feasible to compute. Despite the library being written in C language, it required hundreds of gigabytes of RAM (it was tried in STATA[2] as well).

It was decided to create groups of languages. One can choose Language families or Language groups. Language families would massively decrease the number of dummy variables, but this option might lose much information. Language groups are more granular than the families but still decrease the number of dummy variables significantly. Nine language groups were used to compose the dummy variables.

Table 4.5 informs about the goodness of fit for the hm1_lan model. As expected, the deviance is higher than in the base model. The loglikelihood is worse as well, but it is smaller by a margin.

The estimated effects of the base variables (i.e., Cybersecurity, Password

[2]https://www.stata.com/

length and the Effort) are similar to the base model. Charts and a detailed table with the estimates can be found in the appendix.

Language effect Figure 4.5 and Figure 4.6 reveal the estimated coefficient of the language families (i.e. mutually exclusive dummy variables). Germanic language group was used as the base and thus excluded from the model. The Germanic group includes languages such as German and English.

Figure 4.5: Evaluation of the language dummy variables - Cluster A

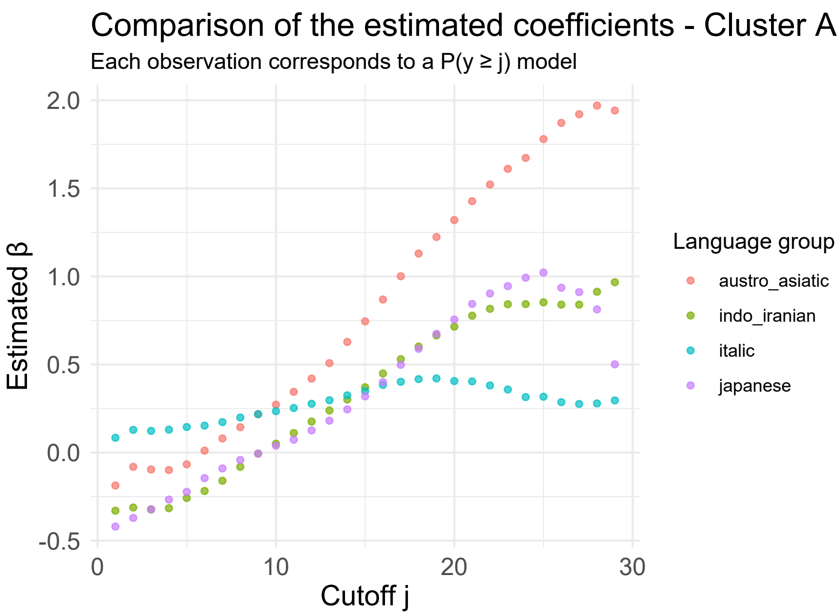


Figure 4.6: Evaluation of the language dummy variables - Cluster B

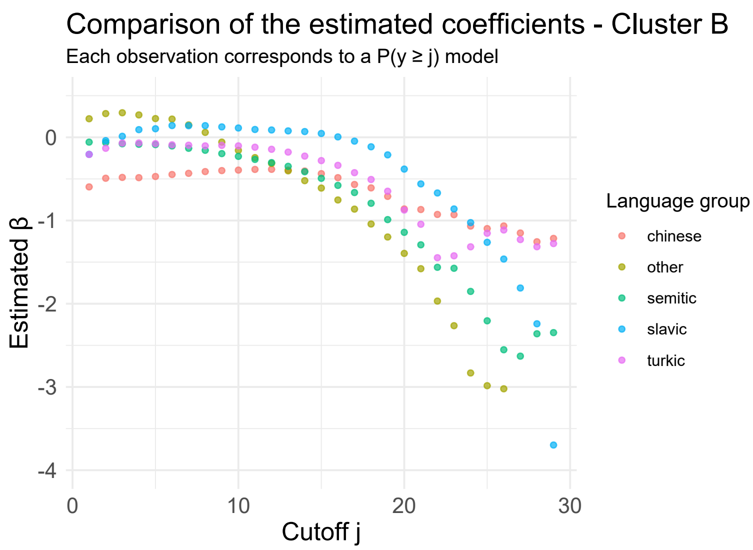


Table 4.6: Significance of language dummy variables

cutoff j	a. asiatic	chinese	i. iranian	italic	japanese	other	semitic	slavic	turkic
2		***	***		***	*		**	
3		***	***	**	***	***			
4		***	***	***	***	***			
5		***	***	***	***	***	*	*	
6		***	***	***	***	***	*	**	
7		***	***	***	*	***	***	***	
8		***	***	***	.	***	***	***	.
9	**	***	**	***		.	***	***	*
10	***	***		***		.	***	***	*
11	***	***	*	***		***	***	***	*
12	***	***	***	***		***	***	***	**
13	***	***	***	***	**	***	***	***	***
14	***	***	***	***	***	***	***	***	***
15	***	***	***	***	***	***	***	**	***
16	***	***	***	***	***	***	***	.	***
17	***	***	***	***	***	***	***		***
18	***	***	***	***	***	***	***		***
19	***	***	***	***	***	***	***	**	***
20	***	***	***	***	***	***	***	***	***
21	***	***	***	***	***	***	***	***	***
22	***	***	***	***	***	***	***	***	***
23	***	***	***	***	***	***	***	***	***
24	***	***	***	***	***	***	***	***	***
25	***	***	***	***	***	***	***	***	***
26	***	**	***	***	***	***	***	***	***
27	***	*	***	**	***	***	***	***	**
28	***	*	***	*	***		***	***	*
29	***	.	***	.	**		***	***	.
30	***		***				*	*	

Furthermore, Table 4.6 informs about the significance of individual estimates. Asterisk indicates whether the estimated β coefficient was statistically different from 0 using the following thresholds of the p-values: 0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 1'. As one can see, not all the coefficients were found to be significant. Detailed standard errors and p-values can be found in the appendix.

Figure 4.5, contains Cluster A of the languages - languages, that seems to have a positive effect on the Password-Username similarity compared with the Germanic group.

Austro-Asiatic group The estimates of the Austro-Asiatic group suggest the most dramatic change within this cluster. The First eight estimated coefficients are statistically indifferent to 0. Nevertheless, the rest of the coefficients is strongly significant and increases through the cutoff j.

One could conclude that there seems to be no difference between the Germanic and Austro-Asiatic group for similar usernames and passwords. However, for moderately to highly dissimilar passwords and usernames, it seems that being from the Austro-Asiatic group leads to more dissimilar passwords and usernames.

Japanese, Indo-Iranian The β estimates of Japanese and Indo-Iranian groups are very similar. The coefficients are strongly significant for highly similar passwords and usernames and highly dissimilar passwords and usernames. The effect seems to be weaker than in the case of the Austro-Asiatic group.

The estimates are negative for the first cutoffs j. That indicates that being from Japanese or Indo-Iranian increases the overall similarity of a password and username for very similar passwords and usernames.

Above cutoff 10, the estimates are positive. That suggests that for moderately to highly dissimilar passwords and usernames, being in Japanese or Indo-Iranian groups contributes positively to the dissimilarity between passwords and usernames.

Italic The estimated coefficients of the Japanese group suggest a relatively flat pattern. The vast majority of the estimates are statistically significant and positive.

The results suggest that speaking an Italic language increases the overall dissimilarity of passwords and usernames. This increase is similar across the dichotomization. On the other hand, it is relatively weak. In the maximum, it would increase the log-odds of being at or above a specific target level by 0.5.

Figure 4.6 shows language groups with negative estimated effects.

Chinese The estimates of the Chinese language group are mostly significantly different from 0 with a weak negative effect. The magnitude does not change much through the cutoff j. The results suggest that speaking a language from the Chinese group increases the overall similarity between a username and a password compared to the Germanic group.

Turkic, Semitic The estimates of the Turkic and Semitic group have a similar course. Negligible effect for highly similar passwords and usernames that increases through the cutoff j. The magnitude for higher cutoff would decrease the log-odds of being at or above a category by more than 2. The estimates

Table 4.7: Goodness of fit measures for hm1_sent model

Model	*Deviance*	*LogLikelihood*	*Degrees of freedom*
hm1_base	685 245.7	-342 622.8	1 310 684
hm1_sent	226 554.7	-113 277.4	481 313

suggest an overall increase in the dissimilarity in comparison with the Germanic group.

Slavic The estimated coefficients of the Slavic group are significant for slightly different passwords and usernames and highly different passwords and usernames. For cutoff 7 to cutoff 14, there is a significant positive but minimal effect on the Password-Username similarity. For high cutoffs, the estimates become significantly strongly negative.

That indicates that for dissimilar passwords and usernames, speaking a Slavic language increases the overall similarity between a username and a password. The practical implications of the middle cutoffs are minimal.

In conclusion, results suggest that the language of the user matters. Some languages contribute positively to the Password-Username similarity and others negatively.

4.1.5 Model family 1 - Sentiment model (m1_sent)

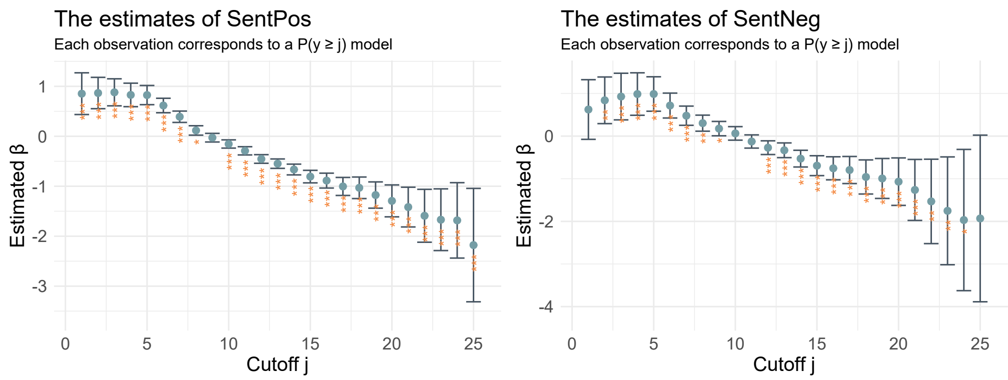
This model aimed to assess whether there are some systematic differences in the Password-Username similarity related to passwords with positive and negative connotations. Two dummy variables extended the base model. One was indicating positive vibes of the password, and the second one, indicating a negative vibes. The base is a neutral password.

Table 4.7 shows the model's basic statistics compared with the base model. The deviance is significantly lower, over 226 thousand, the loglikelihood is higher (-113 thousand) and the degrees of freedom are dramatically lower (481 thousand). These dramatic changes are mostly caused by a lower number of observations used. Not all the passwords were subject to the polarity analysis as described in detail in the Methodology part.

The estimated effects of the base variables (i.e., Cybersecurity, Password length and the Effort) are similar to the base model. Charts and a detailed table with the estimates can be found in the appendix.

Figure 4.7 reveals the β estimates of the sentiment variables in the standard format. Overall, the polarity indicators suggest a significant effect on the Password-Username similarity.

Figure 4.7: The estimated β coefficients of m1_sent model



*Asterisk in a chart indicates whether the estimated β coefficient is statistically different from 0 using following thresholds of the p-values: 0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 1'*

SentPos, SentNeg Interestingly, the estimates of the Positive connotations and Negative connotations dummy variables are very similar. Both are strongly significant for most of the cutoffs j. Furthermore, both groups of estimates are positive for the first third of the estimates and decrease through the cutoff j.

The highest estimate for the Positive dummy variable is nearly one whereas 0.5 for the Negative dummy variable. The lowest estimate for both dummy variables is around -2, corresponding to the highest cutoff j. The chart covers 1 to 26 cutoff. The three last cutoffs were unreasonably low due to the very low number of observations in these bins. This decrease in confidence can be actually seen in most of the course of the estimates.

The results suggest that if a user's password has some connotations (positive or negative), it would affect the similarity between the username and a password. However, this effect is different for similar usernames and passwords and highly dissimilar usernames and passwords.

For very similar passwords and usernames, the polarity of a password contributes to the overall dissimilarity between a username and a password. On the other hand, as the dissimilarity between a username and a password increases, the polarity is related to an increase in the similarity between passwords and usernames.

In conclusion, results suggest that the polarity in the password affects the Password-Username similarity significantly, and this effect is strongly related to the similarity of the password and username.

Table 4.8: Goodness of fit measures for hm1_TLD model

Model	*Deviance*	*LogLikelihood*	*Degrees of freedom*
hm1_base	685 245.7	-342 622.8	1 310 684
hm1_TLD	154 321.4	-71 606.3	258 120

4.1.6 Model family 1 - TLD model (m1_TLD)

This model aimed to assess whether there are some systemic differences in the Password-Username similarity among TLDs (i.e. countries). The initial idea was to create a set of dummy variables, each corresponding to one of the TLDs and observe the effect compared with a base TLD. Unfortunately, it was realized that it is not computationally feasible to do so.

Having more than 150 TLDs in the sample and 30 levels of the target variables, it would be necessary to estimate over five thousands estimates. There would be an issue with the computational capabilities, the number of estimates to interpret and the highly complex parameter space where the optimization should be performed.

To reveal at least some effects, it was decided to make a subset of TLDs, and European countries were selected out of the whole sample. Unfortunately, that would still imply a significant number of dummy variables (i.e. TLDs) to estimate. Consequently, ten countries were chosen for comparison, and the rest was treated as "Others". Similarly to the language-based model, Germany was chosen as the base for the mutually exclusive dummy variables.

The outer categories got even more sparse due to this subsetting. Thus, the maximum of the Password-Username similarity was limited to 25.

Table 4.8 shows the basic fit of the model. Deviance is dramatically lower than in the base model, loglikelihood higher, and as mentioned, degrees of freedom are 258 thousand, dramatically lower than in the base model. But still a large number of observations.

The estimated effects of the base variables (i.e., Password length and the Effort) are similar to the base model. On the contrary, the Cybersecurity index became insignificant. The sampling most probably causes this. European countries have a very similar level of Cybersecurity, and thus, the variable has a very low variance and might not explain much of the Password-Username similarity.

Charts and a detailed table with the estimates can be found in the appendix.

Figure 4.8 and Figure 4.9 shows the estimated coefficient of the TLD dummy variables and Table 4.9 reveals whether the estimated coefficients are significantly different from zero using following thresholds of the p-values: 0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 1'. As one can see, not all the coefficients were found to be significant. Detailed standard errors and p-values can be found in the appendix.

Figure 4.8: Evaluation of the TLD dummy variables - Part A

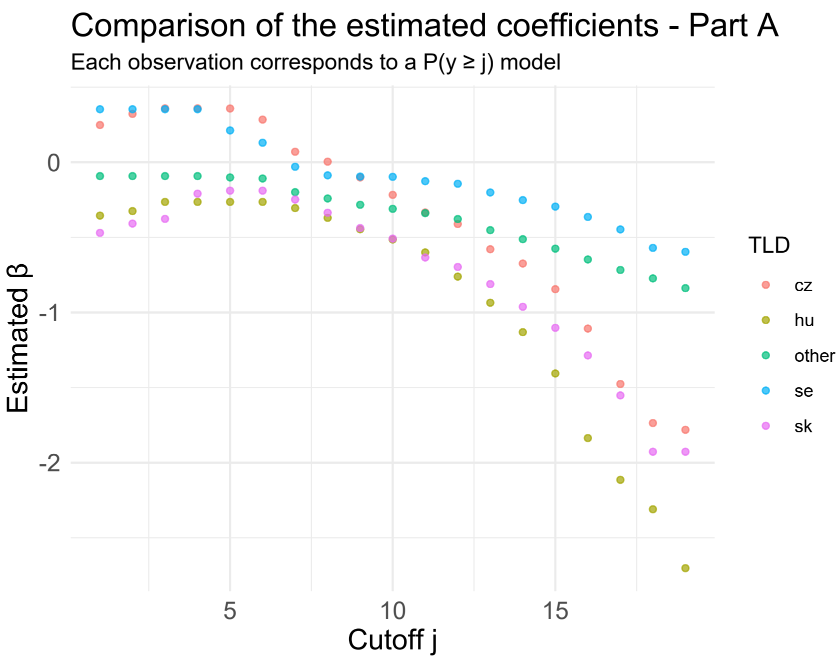


Figure 4.9: Evaluation of the TLD dummy variables - Part B

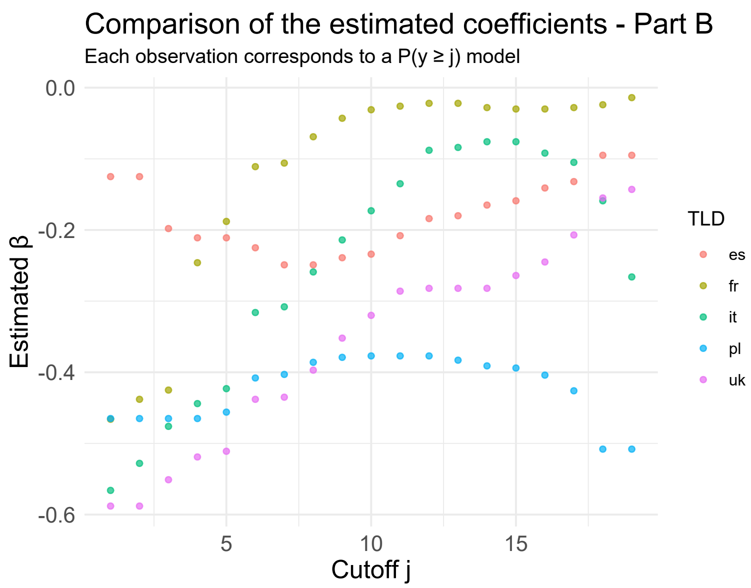


Table 4.9: Significance of TLD dummy variables

cutoff j	cz	es	fr	hu	it	other	pl	se	sk	uk
2			*		**		*		.	**
3	***		***	***	***	***	***	.	***	***
4	***	.	***	***	***	***	***	.	***	***
5	**	.	***	***	***	**	***	.	***	***
6	*	.	**	**	***	*	***		**	***
7	.	*		**	***	*	***		*	***
8		*		***	***	***	***		**	***
9		**		***	***	***	***		***	***
10		**		***	***	***	***		***	***
11	**	**		***	**	***	***	.	***	***
12	***	*		***	*	***	***	*	***	***
13	***	.		***	.	***	***	*	***	***
14	***			***	.	***	***	***	***	***
15	***			***	.	***	***	***	***	***
16	***			***	.	***	***	***	***	**
17	***			***	.	***	***	***	***	**
18	***			***	*	***	***	***	***	.
19	***	.		***	*	***	***	***	***	
20	***	*		***	**	***	***	***	***	
21	***	**		***	*	***	***	***	***	
22	***	**		***	**	***	***	**	***	
23	*	*			*	***	*	*	*	
24		.				***				
25						**				

Figure 4.8 shows the first part of the estimated β coefficients of a group of TLDs with mainly decreasing estimates through the cutoff j.

Czechia The estimates corresponding to Czech users are most significant. For highly similar usernames and passwords, Czech users have fewer similar passwords and usernames than German users (Germany used as the base). However, for dissimilar passwords and usernames, the estimates become negative, indicating that Czech users have more similar passwords and usernames than German users.

Hungary, Slovakia The estimates of Hungarian and Slovakian dummies are very similar. Both are statistically significant and negative for most of the cutoffs. This evidence suggests that Hungarian and Slovak users generally have more similar passwords and usernames than their German colleagues. The difference increases through the cutoff j.

Sweden The estimates of the Swedish indicator are insignificant for the first half of the cutoffs j. However, the second half is statistically significant and negative. That suggests that for dissimilar passwords and usernames, Swedish users have more similar passwords and usernames than German users.

Figure 4.9 describes the rest of the estimates of the selected TLD related dummies. All of the estimates, if significant, suggest a higher similarity of passwords and usernames than German users.

Spain The estimates of the Spanish dummy variable are statistically insignificant to 0 at 5% significance level for all the cutoffs j.

France The estimated coefficients of the French dummy are significant only for three of the cutoffs j. Nevertheless, given the wild course of the estimates through the cutoff j, overall, the variable seems to be insignificant as well.

Italy The estimates of the Italian dummy are significant nearly for the first half of the cutoffs j. Estimated coefficients are negative, suggesting more similar passwords and usernames than in the case of German users. However, the effect is relatively weak, not even -0.5 at the minimum.

Table 4.10: Summary of the Model family 2

	Model (m2_)			
Variable name	m2_full	m2_base	m2_TLD	m2_lan
Cyber	x	x	x	x
Mobile	x			
SexF	x			
TLDs			x	
Languages				x

Poland It seems that Polish users seem to be less responsible than their German counterparts. The estimates of the dummy variable are negative and significant for nearly all of the cutoffs j. Furthermore, the magnitude is more or less constant through the cutoff j (around -0.4).

These results indicate that regardless of the level of the Password-Username similarity, Polish users have overall more similar passwords and usernames than German users.

UK The estimates of the UK dummy variable are negative and significant for the first two-thirds of the cutoffs j. The strength of the effect is decreasing through the cutoff j, and in practice, the effect is rather weak. Overall, UK users have more similar passwords and usernames than German users.

To sum it up, the results suggest significant differences in the Password-Username similarity related to the TLDs. Germany seems to have the least similar passwords and usernames among European countries.

4.2 Model family 2 - the reuse of passwords

As discussed, the goal of Model family 2 is to explain why users reuse passwords, a very dangerous practice.

There are four models of the Model family 2 as outlined in Table 4.10. One can see the connection between a model name suffix (e.g., full or base) and variables included. For example, m2_full indicates Model family 2 and the full version of the model. The terminology is identical to the Model family 1.

Table 4.11: Goodness of fit measures for hm2_full model

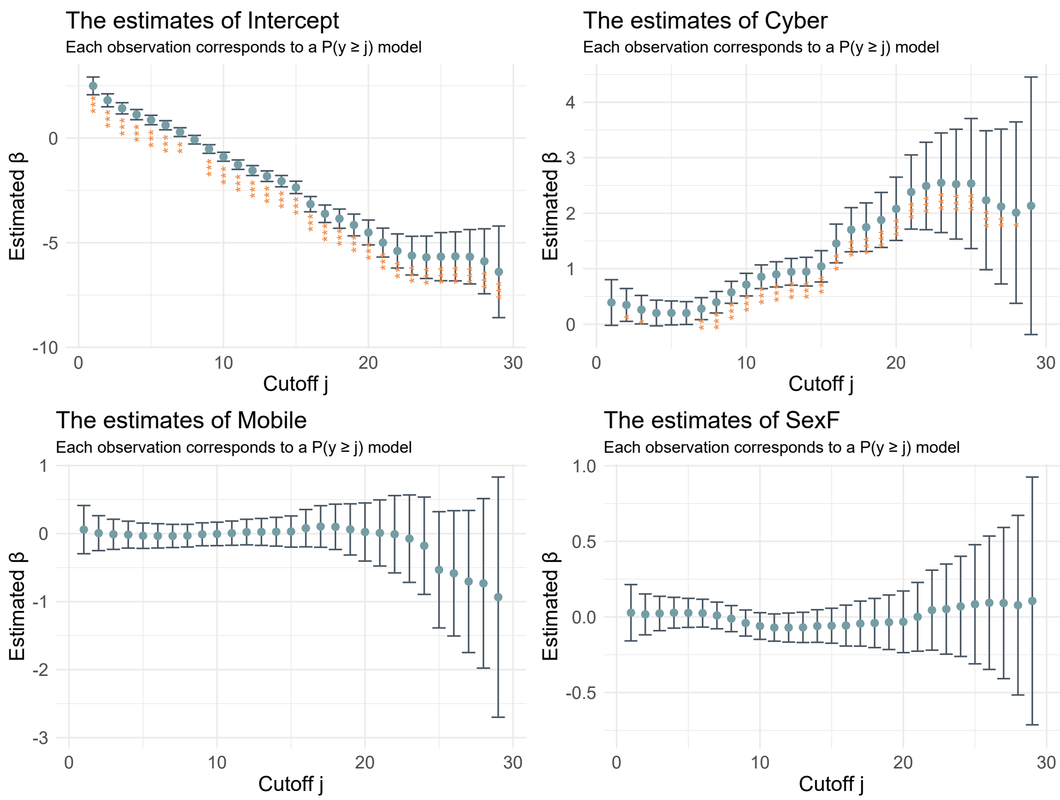
Model	*Deviance*	*LogLikelihood*	*Degrees of freedom*
hm2_full	104492.3	-52246.1	132733

4.2.1 Model family 2 - initial model (m2_full)

This is the initial model of the Model family 2. It contains two macroeconomic variables and the sex. Table 4.11 shows the general statistic of the model. The deviance is over 100 000, the log-likelihood less than 50 000, and the degrees of freedom are above 130 000.

Figure 4.10 shows the estimated effect of the predictors. Overall, the results suggest similar relationships to the Model family 1.

Figure 4.10: The estimated β coefficients of m2_full model



*Asterisk in a chart indicates whether the estimated β coefficient is statistically different from 0 using following thresholds of the p-values: 0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 1'*

Cyber The Cybersecurity index seems to have the most significant effect on the Password-Password similarity.

More than two-thirds of the estimates is significant and greater than zero.

Table 4.12: Goodness of fit measures for hm2_base model

Model	*Deviance*	*LogLikelihood*	*Degrees of freedom*
hm2_full	104 492.3	-52 246.1	132 733
hm2_base	65 213.5	-32 606.77	74 878

A few first and last of the cutoff j estimates are not significant. That might be caused by a lower number of observations in these categories. The effect of the variable increases through the cutoff j, reaching up to 2.5. That indicates that the log-odds of being at or above a category versus below (e.g., for cutoff 25) increases, on average, by 2.5, ceteris paribus.

These findings suggest that Cybersecurity positively affects the dissimilarity of passwords of one user, and this effect increases with the dissimilarity of the passwords.

Mobile The mobile usage seems insignificant for the determination of Password-Password similarity. All the cutoff j estimates are statistically insignificant from 0 at 5% significance level.

Sex The estimated effect of the sex suggest no relationship with the Password-Password similarity. The estimates are insignificant at 5% significance level for all the cutoffs j. The results suggest that the sex does not impact the Password-Password similarity.

4.2.2 Model family 2 - the base model (m2_base)

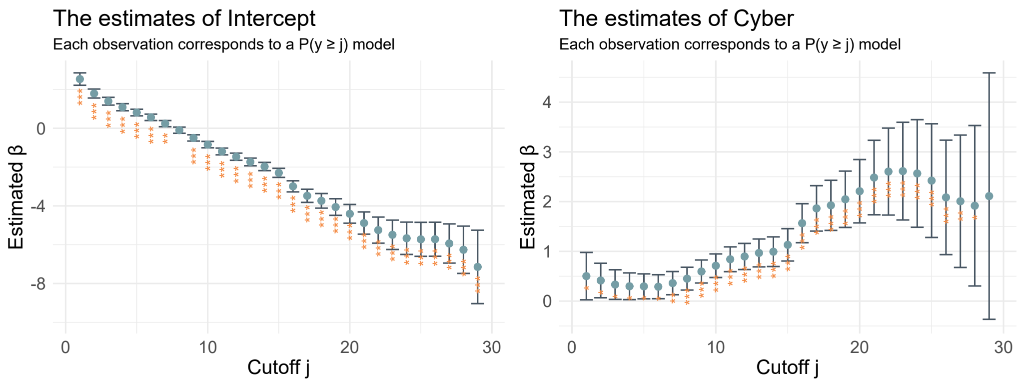
The purpose of this model is to observe what would happen with the initial model if the redundant variables would be omitted. In this case, the Password-Password similarity is being tried to explain by the cybersecurity index.

Table 4.12 reveals the basic statistics of the model. The deviance is two-thirds of the full model. The log-likelihood increased to nearly -32 000, and the degrees of freedom are now almost 75 000.

Figure 4.11 shows the estimated effect of the predictors. In this case, only the cybersecurity index was included.

Cyber The estimated coefficients are strongly significant and positive for most of the cutoffs j. The effect is similar to what has been seen in the full model. Cybersecurity contributes positively to the Password-Password similarity. Users from a country with high Cybersecurity are less prone to recycle passwords.

Figure 4.11: The estimated β coefficients of m2_base model



*Asterisk in a chart indicates whether the estimated β coefficient is statistically different from 0 using following thresholds of the p-values: 0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 1'*

Table 4.13: Goodness of fit measures for hm2_lan model

Model	*Deviance*	*LogLikelihood*	*Degrees of freedom*
hm2_base	65 213.5	-32 606.7	74 878
hm2_lan	62 134.5	-31 067.3	58 896

4.2.3 Model family 2 - language model (m2_lan)

Similarly to the Model family 1, it was expected that language might play an important role in password reuse. Language group dummy variables were added to the base model. The language groups and dummies are identical to the model family 1. As already mentioned, there might not be enough observations in some bins, inducing unrealistic estimates of the cutoffs j. The Password-Password similarity was thus limited to 25.

Table 4.13 reveals the basic goodness of fit measures for the model with language dummy variables. The deviance is slightly lower than in the base model, log-likelihood slightly higher, and as expected, degrees of freedom are lower (the target variable was trimmed).

The estimated effects of the base variables (i.e., Cybersecurity, Password length and the Effort) are similar to the base model. Charts and a detailed table with the estimates can be found in the appendix.

Figure 4.12 and Figure 4.13 informs about the estimated coefficients of the language dummy variables. Furthermore, Table 4.14 reveals whether the coefficients are statistically different from zero. Overall, it seems that the language does not contribute to the reuse of passwords.

While the estimates suggest some structural differences between language groups, Table 4.13 shows that the estimates are statistically indifferent from

Figure 4.12: Evaluation of the language dummy variables - Part A

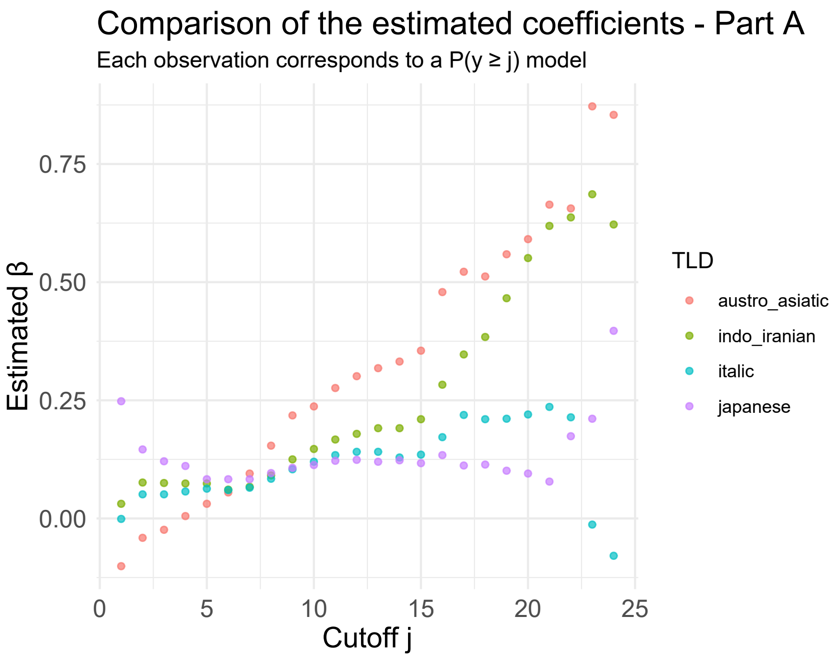


Figure 4.13: Evaluation of the language dummy variables - Part B

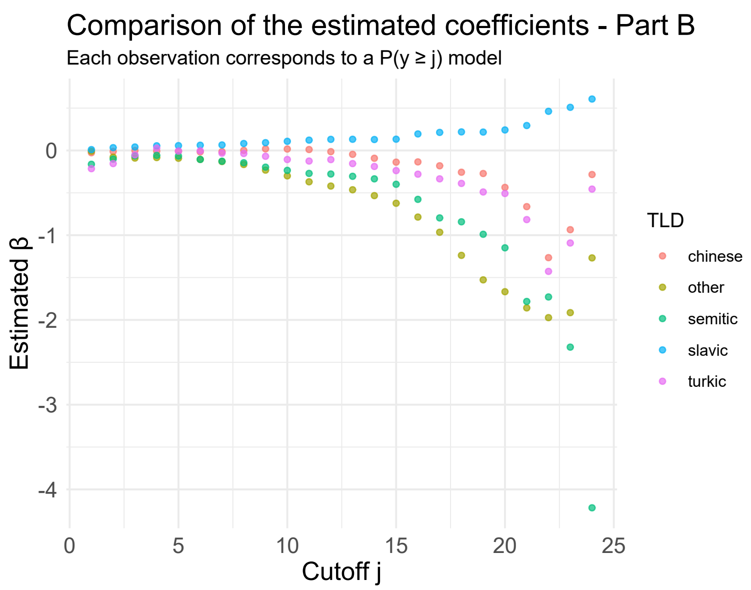


Table 4.14: Significance of the language group dummy variables

language	a. asiatic	chinese	i. iranian	italic	japanese	other	semitic	slavic	turkic
2									
3									
4									
5									
6									
7									
8									
9									
10						*	.		
11						**	*		
12						**	*		
13						**	*		
14						***	*		
15						***	*		
16						***	*		
17						***	**		
18			.			***	**		
19			.			***	**		
20			.			***	**		
21			*			***	**		
22			*			**	**		
23			.			*	*		
24						.	.		
25									

zero. That means that users from the language groups don't have statistically different passwords than Germanic users (base category).

These are exciting findings as the language group did matter for explaining the Password-Username similarity.

4.2.4 Model family 2 - model with domains (m2_TLD)

The Password-Password similarity might be different by countries. This model is an extension of the base model - it includes dummy variables of the TLDs. The dummy variables correspond to the European countries, and Germany was chosen as the base country.

Because of the sparsity of TLDs for high cutoffs j, the dependent variable was capped at 20. Table 4.15 indicates the basic statistics of the model. The deviance is over 56 000, slightly lower than for the base model, log-likelihood is more than -28 000, which is also slightly less than the base model, and finally, there are more almost 42 000 degrees of freedom.

The estimated effects of the base variables (i.e., Cybersecurity, Password length and the Effort) are similar to the base model. Charts and a detailed table with the estimates can be found in the appendix.

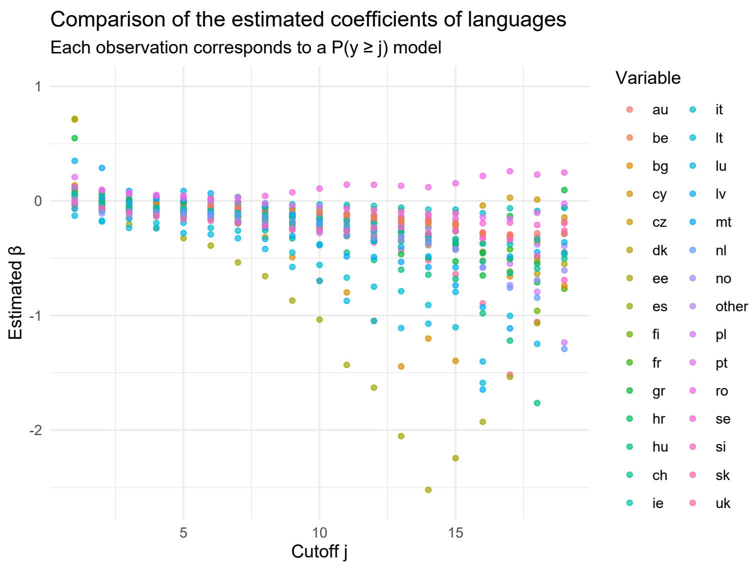
Table 4.15: Goodness of fit measures for hm2_TLD model

Model	*Deviance*	*LogLikelihood*	*Degrees of freedom*
hm2_base	65 213.5	-32 606.7	74 878
hm2_TLD	56 322.7	-28 161.3	41 591

Figure 4.14 gives a high-level overview of the estimated coefficients of the TLD dummy variables. Additionally, Table A.12 informs about the significance of the estimates using following thresholds of the p-values: 0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 1'. As one can see, all the estimated coefficients of the TLD dummy variables are statistically insignificant at 5% significance level. Detailed coefficient estimates can be found in the appendix.

The results suggest no difference in reusing passwords between users from Germany and the selected countries (TLDs).

Figure 4.14: High level evaluation of the TLD dummy variables



Chapter 5

Robustness check

This part aims to check the robustness of the results from two perspectives. First, how the results would change using a different sampling strategy. Second, address the model uncertainty using the Bayesian Model Averaging. Both have been done for the Model family 1, studying the Password-Username similarity, and for the Model Family 2, focusing on reusing of passwords.

5.1 Assessment of the subsampling bias

The available data for this thesis were large. Over 1.4 billion observations. As discussed, it was not feasible to use that many observations mainly for one reasons. The Generalised Ordered Logistic Regression requires very large computational resources. Having 30 different target categories and a few of the independent variables, the optimization algorithm requires hundreds of gigabytes of the RAM.

To mitigate this issue, it was decided to perform the stratified sampling as thoroughly described in the data analysis part. This sampling was dependant on two factors. First, the random seed in R, defining the pseudo randomness and furthermore, the size of the sample.

Two aspects are to be assessed. First, if a different seed would yield similar results and second, if the sample size used for the main sample was enough large.

The strategy was simple. In addition to the sample used for the main results, first, take equally sized sample but generated using a different seed and second, generate a larger sample. The size of the sample is controlled through the base scaling TLD. That is, take more of *.cz* observations which implied larger sample maintaining the population ratios across countries. This was thoroughly described in the data analysis part.

The ordered models were estimated again using these three samples. As a result, one can compare whether the estimated β coefficients vary across the three samples. If there would be minimal differences, one might conclude that the results seem to be robust to the sampling strategy (in terms of size and random seed).

5.1.1 Subsampling bias - Model family 1

This subsection is devoted to the Model family 1, explaining the Password-Username similarity. Figure 5.1 and Figure 5.2 shows the estimated β coefficients of the models based on three different samples.

The individual observations on the charts have the same meaning as in the Results chapter. Each dot is the estimated β coefficient corresponding to one of the cutoffs j (i.e., $P(Y \geq j)$). The color of the dot indicate what sample was used for one of the three models (i.e., model Type). All three models explain the same relationships. Password-Username similarity is being explained by password length, Effort and the cybersecurity. In other words, it is the Base model trained on different samples.

base model, in orange color, is the model presented earlier in the main results. *seed* model, in light gray, is a model trained on equally sized data sampled using a seed 2000. *size* model, in dark gray, is a model based on a sample with seed 3000 and with nearly twice as much observations.

As one can see, there are minor differences between the three samples. For the utmost cutoffs j, there are some differences, but considering the 95% confidence intervals, the vast majority of them do not seem to be significant.

The cutoff 30 estimate of the Effort4 of the base model is significantly lower than the estimates of the other two models. That could be explained by the lower number of observations in that bin, hence the wider confidence intervals.

In conclusion, the results suggest that the sampling strategy (in terms of the size and the random seed) did not significantly affect the estimated coefficients of the base model of the Model family 1.

5.1.2 Subsampling bias - Model family 1

The subsampling bias was investigated for the Model family 2 in a similar way. The base model was estimated using three different samples. A sample used

Figure 5.1: Comparison of the estimated coefficients of the base model trained on different samples - part A

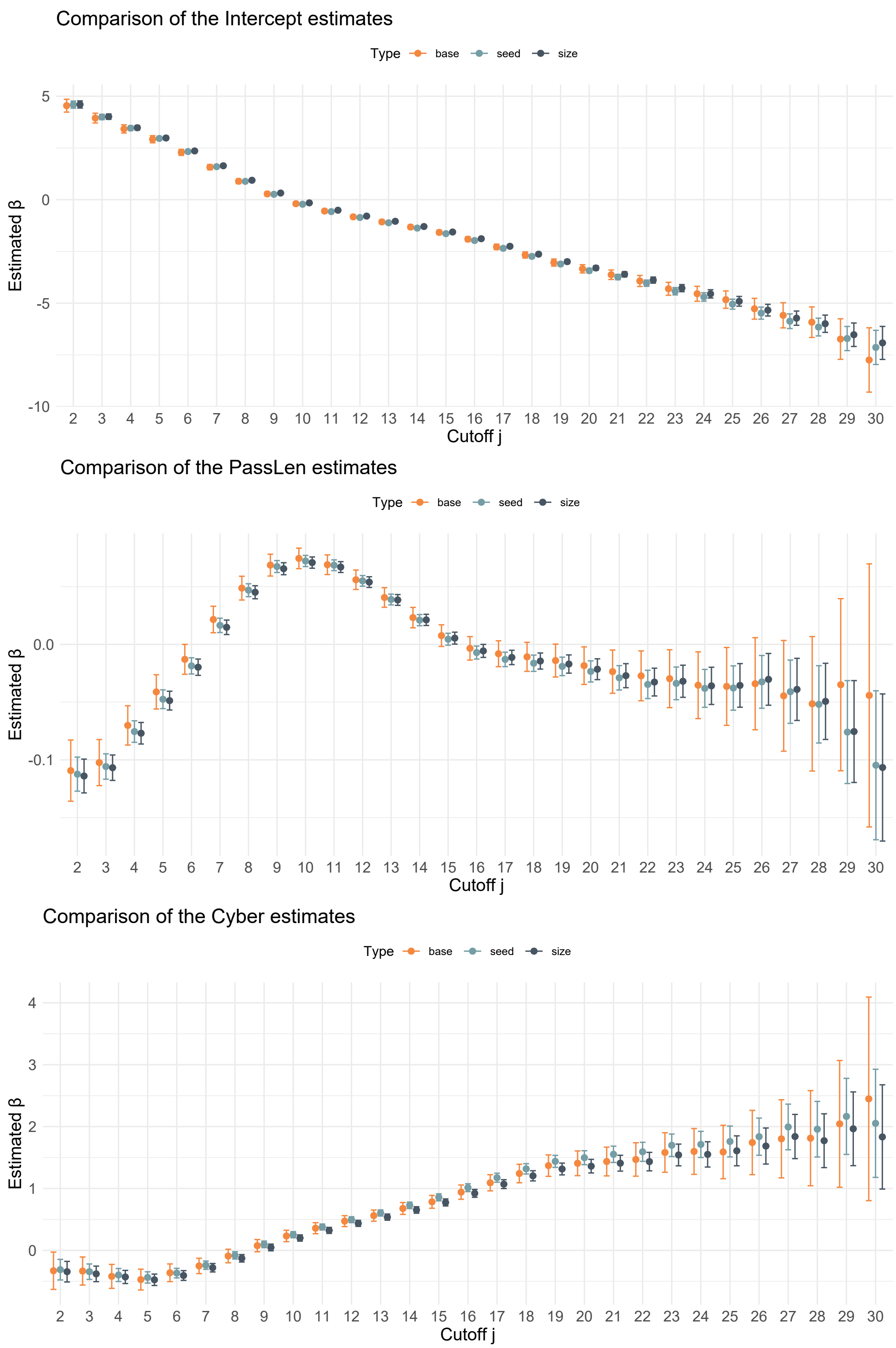
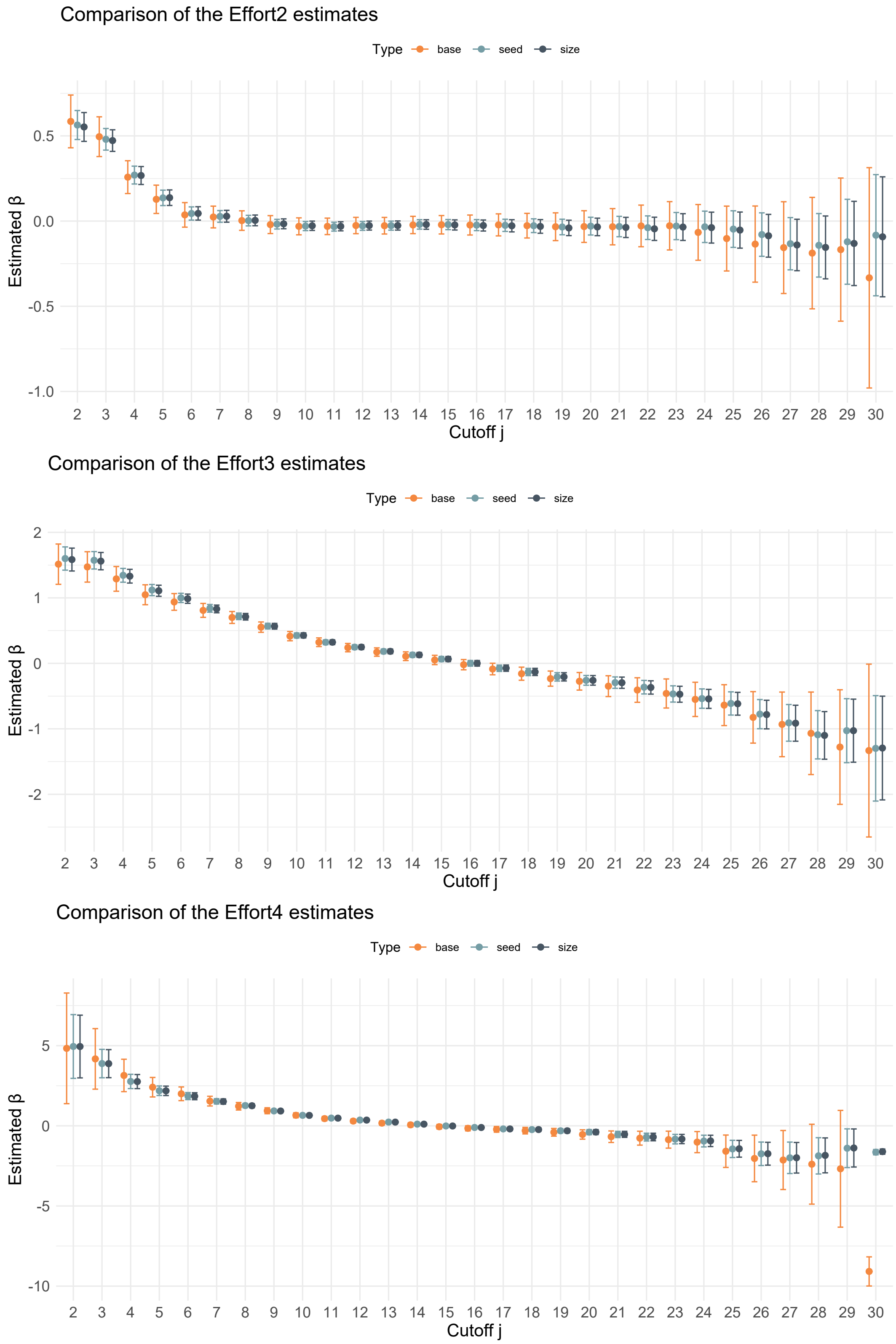


Figure 5.2: Comparison of the estimated coefficients of the base model trained on different samples - part B



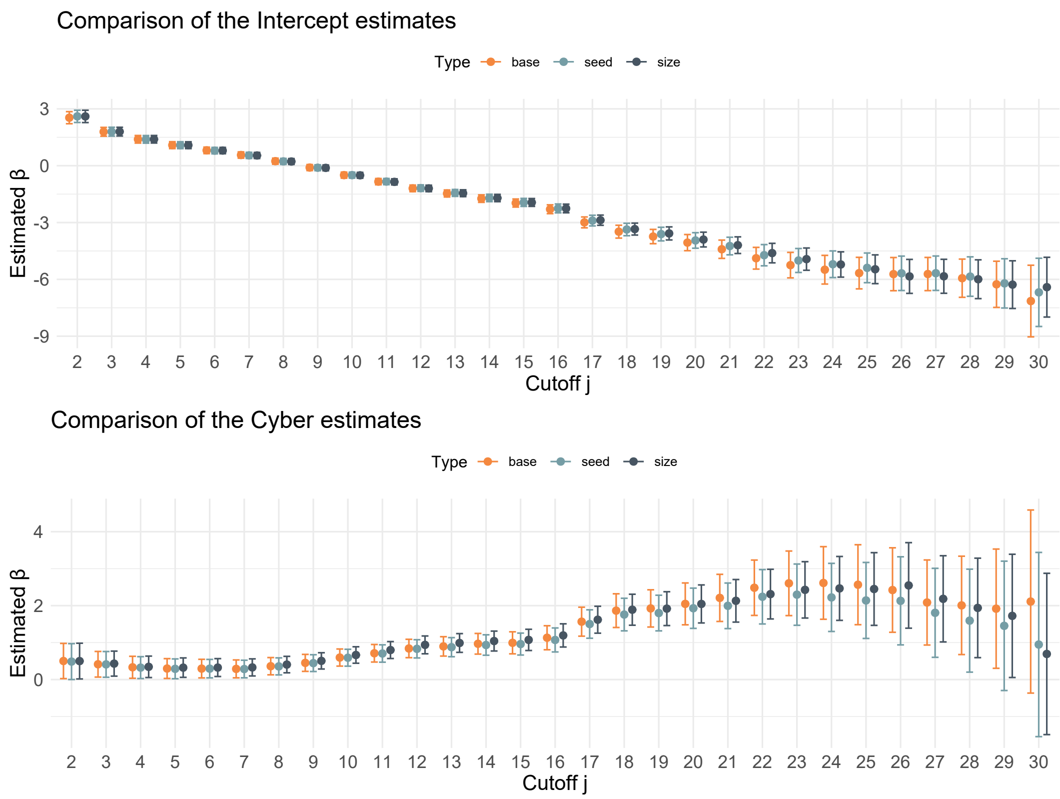
for the main results, a sample with a different seed and a large sample having twice as much observations. The base model for the Model family 2 contains only Cybersecurity index.

Figure 5.3 reveals the estimated β coefficients of the cutoffs j for the three samples (i.e., Type). As before, *base* is the base model presented in the main results, *seed* is a based on a sample generated using a different seed (i.e., 2000) and *size* is a sample generated using the seed 3000 having twice as much observations.

The estimates of the Cybersecurity index are similar across the cutoffs j and the sampling strategies. There are minor differences for the utmost cutoffs j. However, there are large 95% confidence intervals for these estimates.

One could conclude that the sampling strategy does not seem to impact the results significantly.

Figure 5.3: Comparison of the estimated coefficients of the base model trained on different samples - Model family 2



5.2 Bayesian Model Averaging

There was a relatively small amount of variables included in the base models, but the lack of the previous research increased the model specification uncertainty. To address this issue it was decided to employ the Bayesian Model Averaging that should help to identify important variables for the model.

The Bayesian Model Averaging (BMA) allows for an estimation of the probability that a given predictor should be included in the underlying model. For each variable, an essential output of the method is the estimated posterior mean, posterior variance and posterior inclusion probability. The approach estimates 2^n models where n stands for the number of explanatory variables. In other words, it estimates a model using all possible combinations of the variables.

The posterior model probabilities indicate the likelihood of each model. The estimated coefficients weighted using all model instances by the posterior model probabilities are the posterior means. Last, the posterior inclusion probability of a variable is calculated as the sum of posterior model probabilities where the particular variable was retained. Detail description of the BMA procedure provides Raftery *et al.* (1997). A practical example was given by Havranek *et al.* (2020).

The estimation was done in R using the *bms* package[1].

The BMA assumes a continuous variable and in this thesis, the target variable was defined as ordered categorical variable. Thus, the results of the BMA should help to check the robustness on the cost of having imperfectly continuous target. The important outcome of the analysis will be the inclusion of a variable in the optimal model and its sign.

As a consequence of the discussed multicollinearity, the BMA models were estimated using the dilution prior (George 2010) that should account for possible high correlation among the regressors.

5.2.1 BMA of the Model family 1

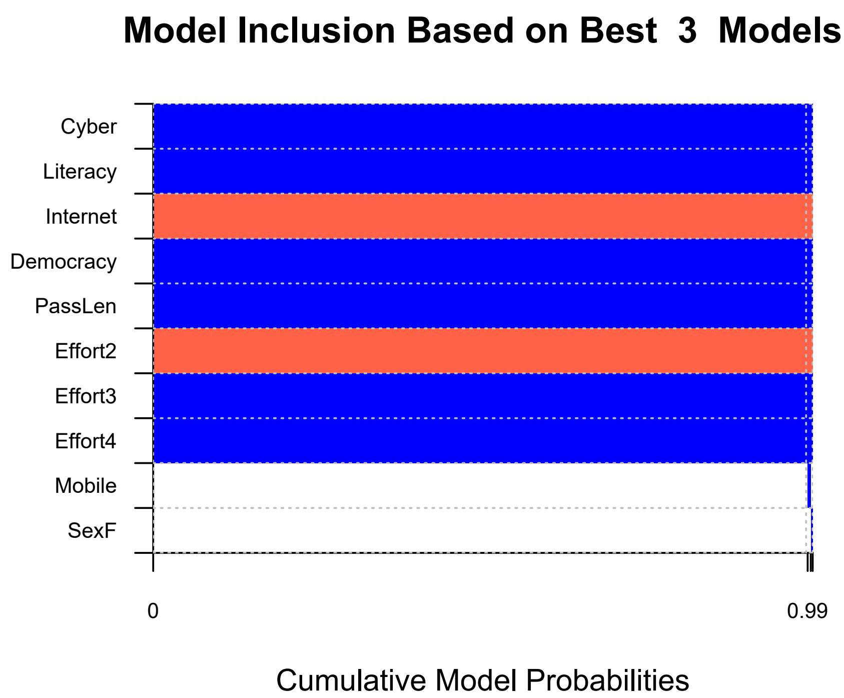
The Bayesian Model Averaging for the Model family 1 was performed using the data set used for the base model described in the Results chapter. This sample had more than 2 million of observations. The Password-Username similarity

[1] https://cran.r-project.org/web/packages/BMS/BMS.pdf

was explained by all five macroeconomic variables, password length, the Effort and gender.

Figure 5.4 informs about the cumulative model probabilities. The best model, plotted on the left-hand side, contains all the macroeconomic variables, password length and the Effort. The posterior probability of this model is 99%. Mobile usage and gender were not included in this best model. One can find details about the estimated coefficients in the appendix in Table A.13.

Figure 5.4: BMA results - variables that could be included in the model explaining the Password-Username similarity



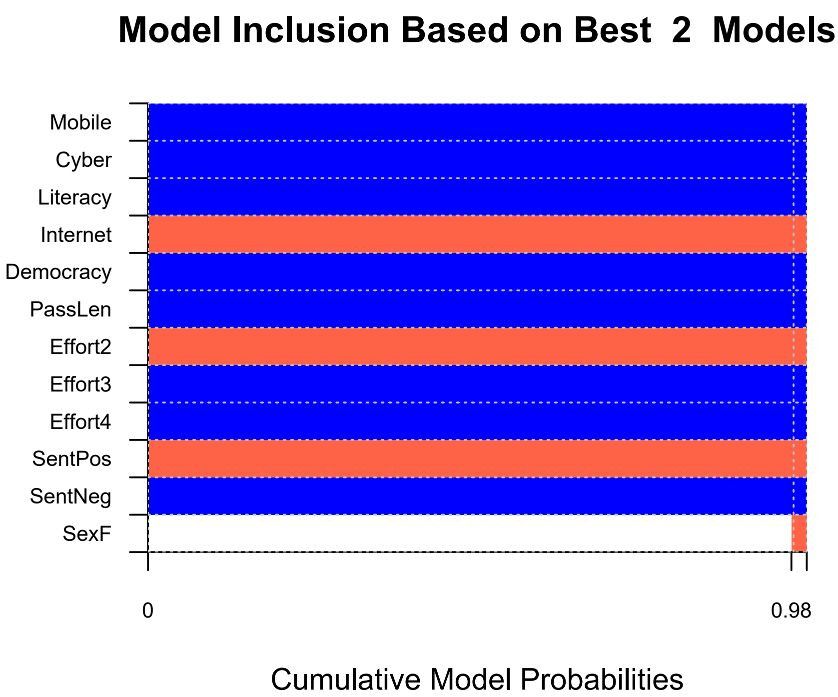
These findings are in line with the observations of the ordered models. Cybersecurity seems to be an essential predictor with a positive effect on the dissimilarity between a password and username. Similarly, the Effort3 and the Effort4 seem to affect the dissimilarity between a username and a password positively. On the other hand, Mobile usage and gender were not included in the model. They were statistically insignificant in the ordered models.

The inclusion of the macroeconomic variables in the best model is in line with the ordered model with the PCA variables. The PCA variables, build on the Literacy rate, Internet usage and Democracy level were statistically significant indicating some joint effect. The BMA suggest that all three variables might be included in the model which is problematic due to the high correlation of the macroeconomic data. That leaves space for further investigation.

In order to assess the importance of the sentiment variables, the BMA has to be performed on a different data. As discussed in the Results chapter, the sentiment was not available for all the languages and thus, the training sample for this kind of a model is smaller.

The BMA for the sentiment data was performed as for the full sample, using the dilution prior alleviating the high correlation among the macroeconomic variables. Figure 5.5 reveals the results of the BMA. As one can see, macroeconomic variables, password length, Effort and the polarity indicators seems to be important. One can find details about the estimated coefficients in the appendix in Table ??.

Figure 5.5: BMA results - variables that could be included in the model explaining the Password-Username similarity including sentiment variables



Both polarity indicators were included in the best model which is in line with the results of the presented ordered model. Both variables were statistically significant for explaining Password-Username similarity.

Contrary to the ordered model, the BMA suggest a positive effect of the positive polarity variable. These finding might be explained by the asymmetrical effects discussed in the ordered logistic regression. The effect of the polarity on the password-username similarity is not identical across the different cutoffs j, which the BMA might not capture.

The Mobile usage was included in the best model with 99% probability. The mobile usage was insignificant in the base ordered model though. The results of the BMS might suggest that the Mobile usage could have some weak effect on the Password-Username similarity that was not captured by the ordered model commented in the Results chapter.

5.2.2 BMA of the Model family 2

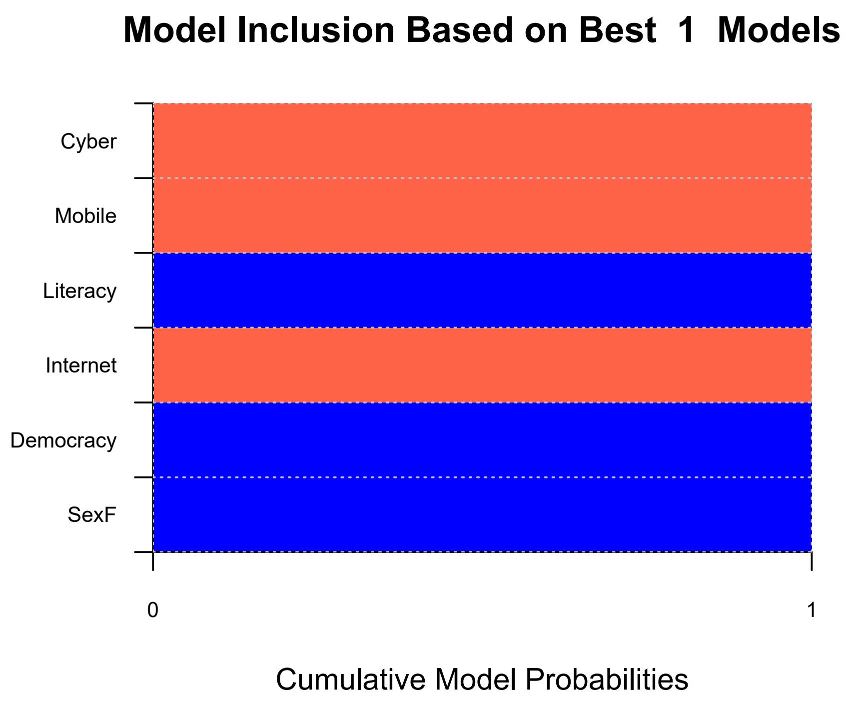
The model uncertainty of the Model family 2 explaining the password reuse was addressed similarly. The BMA was applied on a set of macroeconomic variables and the gender.

Figure 5.6 informs about the calculated cumulative model probabilities. That is, the figure indicates whether a particular variable should be included in the best model. All the macroeconomic variables (i.e., Mobile usage, Cybersecurity, Literacy, Democracy level and Internet usage) were included in the best model. Gender was included in the best model too. One can find details about the estimated coefficients in the appendix in Table A.15.

The inclusion of the gender was unexpected as both ordered models suggested that gender does not play an essential role in explaining the similarity of two passwords. The best model suggested by the BMA, though, would indicate that, overall, females have a lower level of password reuse than males.

Additionally, the best BMA model suggest that the Cybersecurity has a negative effect on the password reuse. However, the estimated ordered model suggested a positive effect on the password reuse. This discrepancy might be caused by the strong correlation among the macroeconomic variables. While the dilution prior helps alleviate the multicollinearity issue, it might not be a bulletproof solution.

Figure 5.6: BMA results - variables that could be included in the model explaining the password reuse



Chapter 6

Discussion

6.1 A comment on the results

The goal of the thesis was to demonstrate that there are differences in password management that some variables could explain. If such differences exist, it would be wise to tailor the password creation policy to individual users. For example, knowing that Czech users are more prone to deriving their passwords from usernames, the provider (e.g., Google) could make an extra effort in convincing Czech users to be careful with their passwords. The main findings of this thesis are following:

1. Higher Cybersecurity level is related to higher password and username dissimilarity
2. Password length significantly affects the password username dissimilarity, but the effect is not monotonous through the cutoff j
3. The character diversity of a password affects the password and username dissimilarity positively for moderately similar passwords and usernames
4. Internet coverage, Literacy rate and Democracy level significantly affect the password username similarity through the PCA, but it needs further investigation.
5. Passwords with positive or negative connotations are related with lower Password-Username similarity for up to moderately similar passwords and usernames

6. The Top Level Domain seems to be associated with the password-username similarity. For example, in comparison with *.de*, *.cz* users have, overall, a higher similarity of passwords and usernames.

7. The language group seems to be associated with the password-username similarity as well. For example, in comparison with the Germanic language group, the Austro-Asiatic language group is related to the higher dissimilarity between passwords and usernames

8. Higher Cybersecurity index is related with lower reuse of passwords

These finding could help to improve the generic password policies. One of the well-known password policies is to use long enough passwords (e.g., at least eight characters) and use various character groups (i.e., lower and upper case letters, numbers and special symbols). However, the fact that the password long does not necessarily mean it is secure. Similarly, using lower and uppercase letters, numbers, and special symbols do not imply a secure password.

The way how people respond to this suggestion is, however, not straightforward. The results of this thesis suggest a couple of drivers of password management that might be leveraged for tailoring the password policies to the users.

The cybersecurity level seems to have a significant effect on both Password-Username similarity and Password-Password similarity. That suggests that the providers (i.e., password policymakers) might make more effort in countries with the lower cybersecurity index to increase the general password management quality.

The results confirm the policy of using a diverse password. While using all four character groups might be perceived as secure, it is also related to lower similarity of passwords and usernames. That is, policymakers should continue emphasising that the diversity of a password is essential.

The passwords policy might also focus on the semantics part of a password. Based on the results, the policymakers might suggest users employ some positive/negative connotations to their passwords as it seems to be related to more dissimilar passwords and usernames. However, that seems to occur for somewhat similar passwords and usernames. If the policymaker knows that the users would choose dissimilar passwords and usernames, suggesting the polarisation might be counterproductive.

The results suggest that the expected country (i.e., TLD) matters. Even though it needs further investigation, if the provider operates on multiple markets, he might ensure the strength of the policy differs among countries. Czech, Hungarian, Slovakian, Polish, and Sweden users have more similar passwords and usernames than German users.

6.2 Potential improvement of the thesis

Given the experimental nature of this thesis and lack of existing literature on this topic, there have been many influential decisions. They were made to the best of the author's knowledge and belief after thorough consideration. This part aims to present potential improvements to what has been done and what might be done differently.

Data One of the strengths of this thesis is the amount of data. Hundreds of millions of observations are not always available in econometric based research. On the other hand, the data are very anonymous, providing little information on the users.

Education and, for example, the digital maturity of a user are unknown, and thus, these variables were used on the national level. It is believed that the user-level granularity of this data would help estimate the effects with higher precision. If, for example, Internet usage was not a significant predictor on the national level, the variation on the user level might contribute significantly to the Password-Username similarity.

Large amount of data with such detailed information are hard to obtain. Nevertheless, it might be attempted to extrapolate this information from smaller data with detailed information about a user. Sex, country, provider and possibly other personal information might be derived from the passwords and usernames that might link the small and detailed data with large and anonymous samples.

Statistical model Based on the empirical results, the Generalised Ordered Regression seems to be a valid approach for modelling selected poor practices of password management. However, one of the disadvantages is the more challenging interpretation and the large number of estimated coefficients. That could be mitigated using the Proportional odds model. That is a Generalised ordered logit that assumes that for a selected variable, the β coefficients are invariant through the cutoff j.

This model would allow for much easier interpretation on the one hand, but on the other, if misused, the asymmetrical effects would be missed. The empirical observation suggests that most of the effects depend on the level of similarity of passwords and usernames.

Dependant variable construction The chosen construction of the predicted variables, the Password-Username similarity and Password-Password similarity, was based on the Levenshtein distance. The interpretation of such a variable is convenient. It is the number of modifications that need to take place to derive one string from the other. During the writing of this thesis, two main questionable properties arose.

First, the Levenshtein distance does not reflect the length of the strings. The practical implications of the distance being 1 are different for strings having 5 and 30 characters. One might assume that for long strings, a short distance is more severe than for short strings. One might consider using the Normalized Levenshtein Distance proposed by Yujian & Bo (2007) that accounts for the length of both strings.

Second, large values of the Levenshtein distance do not indicate which of the strings is accountable for the long distance. For explanatory purposes, it would be helpful to tell whether the long Levenshtein distance is caused by a long username or by a long password.

Polarity assessment The assessment of the polarity of a password is instead a difficult task, mainly due to the lack of training data and very short passwords with special characters. This thesis was based on labelled Twitter data as this data was short enough to be close to passwords and available in multiple languages. Nevertheless, a larger amount of data could improve the accuracy of the polarity detection model. Furthermore, one might try to identify better training data, short enough to match the password structure but universal enough to be found in multiple languages.

Language models The language models used to identify the most probable breakdown of a password into words are highly dependant on the dictionary. Words with typos might construct some passwords. On the contrary, the dictionary for the language models might contain typos that do not appear in the passwords. Both might eventually affect the quality of password parsing.

Thus, it might be considered to build a model for typos correction that would amend the text.

Derivation of the data Some of the predictors, such as gender, had to be derived from the data. This derivation is difficult and time-consuming, but there are not many other options with this kind of data.

In terms of gender, there might be a clear improvement. It was managed to identify the gender for nearly 50% of the users. Three elementary reasons might cause the failure of sex detection. First, names apply to both genders; second, usernames do not contain a given name; and third, usernames contain names in a non-standard format. It might be possible to get dictionaries of first names, including multiple variants. That would help to detect a higher percentage of users.

Additionally, there might be an alternative way of gender detection. For example, in the Czech Republic, the female surname usually ends with *"ová"* suffix. Similar rules of different languages might reveal more of the gender.

Passwords based on keyboard patterns are another example of a bad password management practice. The *qwerty* password is well known. There is no reason to believe that the patterns would affect the Password-Username similarity, but it might affect the Password-Password similarity significantly. It would be feasible to synthesise many similar keyboard patterns and use an indicator of such a pattern in the model.

Age would be an interesting variable to have. It is not directly present in the data, but it might be possible to derive it with some uncertainty. A few digits frequently appear in the username, especially as a suffix. The digits might represent the year of birth. 1965, 1980 or 1991, all these numbers might be considered as birth years. On the contrary, passwords might not be helpful for age detection. If such a number would appear in the password, it might be the birth year of a spouse or child.

There might be special categories of words in the passwords. Consider famous cities, month names, movies and other categories. It might be reasonable to expect that such well-known words would be prone to be shared across multiple passwords of one user. It would be feasible to build such a group of words for one language but having dozens of languages in the sample makes the analysis very difficult.

Another topic for further investigation are special structural patterns of passwords. For example, if a user is forced to change a password, one might append a digit to the tail of the password. A user might be incrementing this digit as the system periodically asks to change the password. This practice is dangerous as a password might be easily derived from an older one.

TLDs, Language groups The results suggest differences in password management among the TLDs. However, it was managed to use only a few of them. It might be helpful to derive a methodology that would allow observing the differences caused by the TLD on a large scale. Policymakers could then focus on these markets in order to improve password management.

Similarly, the language group seems to be related to password management as well. Unfortunately, one language group contain multiple languages. For example, the Germanic language group contains Gothic, Danish, Swedish, Norwegian, Faroese, Icelandic, Bavarian, German, Luxembourgish, Schwytzertütsch, Walser, Yiddish, Afrikaans, Dutch, Flemish, Saxon, English and Frisian language. The individual languages might have a different effect on password management, and the grouped dummy variable might hide such a piece of information.

Chapter 7

Conclusion

This thesis aims to explain two examples of poor password management. First, why users use similar passwords and a username and second, why they reuse their passwords. Both practices present a security threat to the user's data, and the main drivers of such behaviour are unknown. The findings might be used for better password policies increasing the overall security of the user's data.

One of the main results is related to how well a country is resilient to digital threats. This state is often understood as cybersecurity. It might be reasonable to believe that cybersecurity also affects user's behaviour. One of the main results of this thesis is that cybersecurity, measured as the Global Security Index, contributes to a lower similarity between a username and a password. Users living in countries with high cybersecurity might be well aware of the potential digital risks and might want to protect their data accordingly.

Another main result is related to the existing password policy. A typical policy suggests using various characters in a password (i.e., letters, numbers, symbols). It makes the password harder to guess. The results of this thesis shows that this practice has further implications. It seems that including all the character types in a password decreases the similarity of a username and a password. That means that providers (e.g., Google) should continue enforcing these rules.

Cybersecurity is not the only macroeconomic factor affecting users. Digitisation, education and freedom might also be related to people's security awareness. The results suggest that these three factors have a joint effect on the similarity of passwords and usernames. Users in countries with high digital literacy might be well aware of the cyber threats and react accordingly. Similarly, educated people could be well informed about the potential digital risks and choose their password wisely. Users in countries struggling with human rights might want to protect their data more than users in safe countries.

The digitisation was expressed as the internet coverage, education as the literacy rate and freedom was measured using the democracy level. The effects of these variables need further investigation.

The password management was also studied from the cultural perspective. As existing surveys suggests, there might be differences among cultures and languages. This thesis investigated the effect of a country and language group on the similarity between a username and password. As expected, it was found that there are fixed effects among countries and language groups as well. For example, Czech users have, overall, more similar passwords than German users. Similarly, users from the Slavic language group have more similar passwords and usernames than their colleagues from the Germanic language group.

Besides the Password-Username similarity, this thesis also studies password reuse. That is, what makes people derive their passwords on historical ones. It was aimed to explain this practice mainly by macroeconomic variables as in the case of Password-Username similarity. As before, the cybersecurity level might affect users security practices. The results confirm this idea in this case too. It has been shown that users living in countries with high cybersecurity are less prone to reuse their passwords.

The findings might contribute to more robust data security. The identified drivers could be used by providers (e.g., Google) to tailor the password policies to the users. Currently, the recommendations on the password properties are frequently limited to general rules only. For example, Providers could focus on countries with a low Cybersecurity index where the users are likely to have poor passwords and convince them to pay more attention to their decision making.

Similarly, the providers should continue to encourage users to use diverse passwords. The character diversity makes the password stronger, and furthermore, it appears that it leads to a lower similarity between a username and a password.

Researchers might further study the user's behaviour from various perspective. For example, investigate additional predictors, improve the definition of target variables, focus on the detail of password data, elaborate thoroughly on the sentiment detection and its effect or detailed design of password policy. Nevertheless, three main points are following.

First, the findings suggest some effect of the Democracy level, Literacy rate and Internet coverage which are highly correlated. Solving the multicollinearity issue and estimating the effects might bring valuable findings for the tailored password policies.

Second, the results suggest differences among countries. Unfortunately, the large number of countries was not feasible to process. Estimating every Top Level Domain's fixed effect would allow for targeted and tailored password policies to the countries with the worst user's attitude.

Last, an advantage of this thesis is a large amount of data. On the contrary, there is little detailed information about the users. It might be suggested to derive more user-specific information (e.g., age) that would help to explain password management issues as the existing knowledge suggests.

The thesis gives an example of how to study two examples of poor password management. It demonstrates how researchers could use password data to study such a problem and suggest a statistical model to measure the relationship. It also shows the effect on the similarities of a few variables. Furthermore, it describes how to estimate the word composition of a password and model its polarity. Last, it discus what implications the findings might have on the password policies.

Bibliography

A., V. & S. Sonawane (2016): "Sentiment Analysis of Twitter Data: A Survey of Techniques." *International Journal of Computer Applications* **139(11)**.

Alkaldi, N. & K. Renaud (2016): "Why do people adopt, or reject, smartphone password managers?" In "Proceedings 1st European Workshop on Usable Security," Internet Society.

Bakshi, R. K., N. Kaur, R. Kaur, & G. Kaur (2016): "Opinion mining and sentiment analysis." In "Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016," pp. 452–455. Institute of Electrical and Electronics Engineers Inc.

Belenko, A. & D. Sklyarov (2012): "„Secure Password Managers "and „Military-Grade Encryption "on Smartphones: Oh, Really." *Blackhat Europe* pp. 1–12.

Blocki, J., B. Harsha, & S. Zhou (2018): "On the Economics of Offline Password Cracking." In "Proceedings - IEEE Symposium on Security and Privacy," volume 2018-May, pp. 853–871.

Bonneau, J. (2012): "The science of guessing: Analyzing an anonymized corpus of 70 million passwords." *Proc. of Oakland* pp. 538–552.

Bonneau, J., C. Herley, P. C. Van Oorschot, & F. Stajano (2012): "The quest to replace passwords: A framework for comparative evaluation of web authentication schemes." In "Proceedings - IEEE Symposium on Security and Privacy," pp. 553–567. Institute of Electrical and Electronics Engineers Inc.

Bonneau, J. & E. Shutova (2012): "Linguistic properties of multi-word passphrases." In J. Blyth, S. Dietrich, & L. J. Camp (editors), "Financial

Cryptography and Data Security," pp. 1–12. Berlin, Heidelberg: Springer Berlin Heidelberg.

Boser, B. E., I. M. Guyon, & V. N. Vapnik (1992): "A training algorithm for optimal margin classifiers." In "Proceedings of the fifth annual workshop on Computational learning theory," pp. 144–152.

Boucher, J. & C. E. Osgood (1969): "The pollyanna hypothesis." *Journal of Verbal Learning and Verbal Behavior* **8(1)**: pp. 1 – 8.

Bošnjak, L. & B. Brumen (2019): "Rejecting the death of passwords: Advice for the future." *Computer Science and Information Systems* **16**: pp. 313–332.

Breiman, L. (2001): "Random forests." *Machine Learning* **45(1)**: p. 5–32.

Brenner, P. S. & J. DeLamater (2016): "Lies, damned lies, and survey self-reports? identity as a cause of measurement bias." *Social Psychology Quarterly* **79**.

Brown, A. S., E. Bracken, S. Zoccoli, & K. Douglas (2004): "Generating and remembering passwords." *Applied Cognitive Psychology* **18(6)**: pp. 641–651.

Bulbulia, Z. & M. Maharaj (2013): "Factors that influence young adults' online security awareness in the Durban region of South Africa." *Journal of Information Warfare* **12(1)**: pp. 83–96.

Chen, T., R. Xu, Y. He, & X. Wang (2017): "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN." *Expert Systems with Applications* **72**.

Chen, Y. & S. Skiena (2014): "Building sentiment lexicons for all major languages." In "Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)," Association for Computational Linguistics.

Chiasson, S., P. C. van Oorschot, & R. Biddle (2006): "A usability study and critique of two password managers." In "Proceedings of the 15th Conference on USENIX Security Symposium - Volume 15," USENIX-SS'06. Berkeley, CA, USA: USENIX Association.

Cox, D. R. (1958): "The regression analysis of binary sequences." *Journal of the Royal Statistical Society. Series B (Methodological)* **20(2)**: pp. 215–242.

Cravo, C. V., G. Gallegos, B. Castro Armas, D. D'Elia, S. Droz, L. Dandurand, M. Alsalamin, J. Yun, A. Al-Rashidi, N. Radoja, M. Vojvodic, S. Xu, K. L. Lee, M. Humeau, J. Crisp, D. Susar, F. Bosco, & M. Musumeci (2019): *Global Cybersecurity Index 2018*. Geneva, Switzerland: International Telecommunication Union, first edition edition.

Das, A., J. Bonneau, M. Caesar, N. Borisov, & X. Wang (2014): "The tangled web of password reuse." In "NDSS," NDSS Symposium 2014.

Davidov, D., O. Tsur, & A. Rappoport (2010): "Enhanced sentiment learning using twitter hashtags and smileys." In "Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference," volume 2, pp. 241–249.

Fang, Y., K. Liu, F. Jing, & Z. Zuo (2019): "Password guessing based on semantic analysis and neural networks." In "Communications in Computer and Information Science," volume 960, pp. 84–98. Springer Verlag.

Fathy, M. E., V. M. Patel, & R. Chellappa (2015): "Face-based active authentication on mobile devices." In "2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)," pp. 1687–1691.

Filatova, E. (2012): "Irony and sarcasm: Corpus generation and analysis using crowdsourcing." In "Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012," pp. 392–398. European Language Resources Association (ELRA).

Florencio, D. & C. Herley (2007): "A large-scale study of web password habits." In "Proceedings of the 16th International Conference on World Wide Web," WWW '07, pp. 657–666. New York, NY, USA: ACM.

Fullerton, A. S. & J. C. Dixon (2010): "Generational conflict or methodological artifact? reconsidering the relationship between age and policy attitudes in the u.s., 1984-2008." *The Public Opinion Quarterly* **74(4)**: pp. 643–673.

GARCIA, D., A. GARAS, & F. SCHWEITZER (2012): "Positive words carry less information than negative words." *EPJ Data Science* **1(1)**: p. 3.

GEORGE, E. I. (2010): *Dilution priors: Compensating for model space redundancy*, p. 158–165. Institute of Mathematical Statistics.

GHIASSI, M., J. SKINNER, & D. ZIMBRA (2013): "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network." *Expert Systems with Applications* **40(16)**: pp. 6266–6282.

GIACHANOU, A. & F. CRESTANI (2016): "Like it or not." *ACM Computing Surveys* **49(2)**: p. 1–41.

GOUR, S. (2014): "Cyber attacks : An impact on economy to an organization." *International Journal of Information Computation Technology* **4**.

HANAMSAGAR, A., S. S. WOO, C. KANICH, & J. MIRKOVIC (2016): "How users choose and reuse passwords." *University of Southern California* .

HAQUE, S. M., M. WRIGHT, & S. SCIELZO (2014): "Hierarchy of users' web passwords: Perceptions, practices and susceptibilities." *International Journal of Human Computer Studies* **72**.

HAQUE, S. T., M. WRIGHT, & S. SCIELZO (2013): "A study of user password strategy for multiple accounts." In "Proceedings of the Third ACM Conference on Data and Application Security and Privacy," CODASPY '13, pp. 173–176. New York, NY, USA: ACM.

HAVRANEK, T., Z. IRSOVA, L. LASLOPOVA, & O. ZEYNALOVA (2020): "Skilled and Unskilled Labor Are Less Substitutable than Commonly Thought." *Technical report*.

HEAFIELD, K. (2011): "KenLM: faster and smaller language model queries." In "Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation," pp. 187–197. Edinburgh, Scotland, United Kingdom.

HEEN, O. & C. NEUMANN (2017): "On the privacy impacts of publicly leaked password databases." In M. POLYCHRONAKIS & M. MEIER (editors), "Detection of Intrusions and Malware, and Vulnerability Assessment," pp. 347–365. Cham: Springer International Publishing.

Helkala, K. & T. H. Bakås (2014): "Extended results of norwegian password security survey." *Information Management and Computer Security* **22**.

Houshmand, S., S. Aggarwal, & R. Flood (2015): "Next gen pcfg password cracking." *IEEE Transactions on Information Forensics and Security* **10(8)**: pp. 1776–1791.

Jakobsson, M. & M. Dhiman (2013): *The Benefits of Understanding Passwords*, pp. 5–24. New York, NY: Springer New York.

Jianqiang, Z., G. Xiaolin, & Z. Xuejun (2018): "Deep convolution neural networks for twitter sentiment analysis." *IEEE Access* **6**: p. 23253–23260.

Jurafsky, D. & J. H. Martin (2019): "N-gram language models." *Speech and Language Processing* **2019(1)**: p. Chapter 3.

Kamp, P. H. (2012): "LinkedIn password leak: Salt their hide." *Queue* **10(6)**: p. 20.

Kelley, P., S. Komanduri, M. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. Cranor, & J. Lopez (2012): "Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms." pp. 523–537.

Komanduri, S., R. Shay, P. Kelley, M. Mazurek, L. Bauer, N. Christin, L. Cranor, & S. Egelman (2011a): "Of passwords and people: Measuring the effect of password-composition policies." pp. 2595–2604.

Komanduri, S., R. Shay, P. Kelley, M. Mazurek, L. Bauer, N. Christin, L. Cranor, & S. Egelman (2011b): "Of passwords and people: Measuring the effect of password-composition policies." pp. 2595–2604.

Kontaxis, G., E. Athanasopoulos, G. Portokalidis, & A. D. Keromytis (2013): "SAuth: Protecting user accounts from password database leaks." In "Proceedings of the ACM Conference on Computer and Communications Security," pp. 187–198.

Kontopoulos, E., C. Berberidis, T. Dergiades, & N. Bassiliades (2013): "Ontology-based sentiment analysis of twitter posts." *Expert Systems with Applications* .

Li, Y., H. Wang, & K. Sun (2016): "A study of personal information in human-chosen passwords and its security implications." In "Proceedings - IEEE INFOCOM," volume 2016-July. Institute of Electrical and Electronics Engineers Inc.

Li, Z., W. He, D. Akhawe, & D. Song (2014): "The emperor's new password manager: Security analysis of web-based password managers." In "23rd USENIX Security Symposium (USENIX Security 14)," pp. 465–479. San Diego, CA: USENIX Association.

Long, J. & J. Freese (2014): *Regression models for categorical dependent variables using stata*. Stata Press, 3rd edition edition.

Lyastani, S. G., M. Schilling, S. Fahl, M. Backes, & S. Bugiel (2018): "Better Managed Than Memorized? Studying the Impact of Managers on Password Strength and Reuse." In "Proc. USENIX Security Symposium," pp. 203–220.

Ma, J., W. Yang, M. Luo, & N. Li (2014): "A study of probabilistic password models." In "Proceedings - IEEE Symposium on Security and Privacy," pp. 689–704. Institute of Electrical and Electronics Engineers Inc.

Maddux, J. & R. Rogers (1983): "Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change." *Journal of Experimental Social Psychology* **19**: pp. 469–479.

Malone, D. & K. Maher (2012): "Investigating the distribution of password choices." In "WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web," pp. 301–310.

Martínez-Cámara, E., M. T. Martín-Valdivia, L. A. Ureña-López, & A. R. Montejo-Ráez (2014): "Sentiment analysis in Twitter." *Natural Language Engineering* **20(1)**.

Mathur, S., A. Vjay, J. Shah, S. Das, & A. Malla (2016): "Methodology for partial fingerprint enrollment and authentication on mobile devices." In "2016 International Conference on Biometrics (ICB)," pp. 1–8.

Mazumdar, J. (2018): "Retina based biometric authentication system: A review." *International Journal of Advanced Research in Computer Science* **9**: pp. 711–718.

McCullagh, P. (1980): "Regression models for ordinal data." *Journal of the Royal Statistical Society. Series B (Methodological)* **42(2)**: pp. 109–142.

McGill, T. & N. Thompson (2018): *Gender Differences in Information Security Perceptions and Behaviour.* University of Technology, Sydney.

Minsky, M. (1961): "Steps toward artificial intelligence." *Proceedings of the IRE* **49(1)**: pp. 8–30.

Mittal, A. & A. Goel (2012): "Stock prediction using twitter sentiment analysis." *Stanford University* .

Moore, D. H. (1987): "Classification and regression trees, by leo breiman, jerome h. friedman, richard a. olshen, and charles j. stone. brooks/cole publishing, monterey, 1984,358 pages." *Cytometry* **8(5)**: p. 534–535.

Moore, T. (2010): "The economics of cybersecurity: Principles and policy options." *International Journal of Critical Infrastructure Protection* **3**.

Mozetič, I., M. Grčar, & J. Smailović (2016): "Multilingual twitter sentiment classification: The role of human annotators." *PLOS ONE* **11(5)**: p. e0155036.

Mr. S. M. Vohra, P. J. B. T. (2012): "A Comparative Study Of Sentiment Analysis Techniques." *Journal Of Information, Knowledge And Research In Computer Engineering* **(October)**: pp. 313–317.

Mäntylä, M. V., D. Graziotin, & M. Kuutila (2018): "The evolution of sentiment analysis—a review of research topics, venues, and top cited papers." *Computer Science Review* **27**: pp. 16 – 32.

Notoatmodjo, G. & C. Thomborson (2009): "Passwords and perceptions." *Conferences in Research and Practice in Information Technology Series* **98**: pp. 71–78.

Ogbanufe, O. & D. J. Kim (2018): "Comparing fingerprint-based biometrics authentication versus traditional authentication methods for e-payment." *Decision Support Systems* **106**: pp. 1 – 14.

Petrie, H. & B. Merdenyan (2016): "Cultural and gender differences in password behaviors: Evidence from china, turkey and the uk." In "Proceedings of the 9th Nordic Conference on Human-Computer Interaction," NordiCHI '16, pp. 9:1–9:10. New York, NY, USA: ACM.

Pilar, D. R., A. Jaeger, C. F. A. Gomes, & L. M. Stein (2012): "Passwords usage and human memory limitations: A survey across age and educational background." *PLOS ONE* **7(12)**: pp. 1–7.

Raftery, A. E., D. Madigan, & J. A. Hoeting (1997): "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association* **92(437)**: pp. 179–191.

Rao, A., B. Jha, & G. Kini (2013): "Effect of grammar on security of long passwords." In "Proceedings of the Third ACM Conference on Data and Application Security and Privacy," CODASPY '13, pp. 317–324. New York, NY, USA: ACM.

Riddle, B. L., M. S. Miron, & J. A. Semo (1989): "Passwords in use in a university timesharing environment." *Computers and Security* **8(7)**: pp. 569–579.

Saif, H., Y. He, & H. Alani (2012): "Semantic sentiment analysis of twitter." In "The Semantic Web – ISWC 2012," pp. 508–524. Berlin, Heidelberg: Springer Berlin Heidelberg.

Shay, R., S. Komanduri, A. L. Durity, P. Huh, M. L. Mazurek, S. M. Segreti, B. Ur, L. Bauer, N. Christin, & L. F. Cranor (2016): "Designing password policies for strength and usability." *ACM Transactions on Information and System Security* **18(4)**.

Shelar, A. & C. Y. Huang (2018): "Sentiment analysis of twitter data." In "Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018," .

Shen, C., T. Yu, H. Xu, G. Yang, & X. Guan (2016): "User practice in password security: An empirical study of real-life passwords in the wild." *Computers and Security* **61**: pp. 130–141.

Shinzaki, T. (2020): *Use Case of Palm Vein Authentication*, pp. 145–158. Cham: Springer International Publishing.

Stobert, E. & R. Biddle (2018): "The password life cycle." *ACM Trans. Priv. Secur.* **21(3)**: pp. 13:1–13:32.

Suppala, K. & N. Rao (2019): "Sentiment analysis using naïve bayes classifier." *International Journal of Innovative Technology and Exploring Engineering* **8(8)**: pp. 264–269.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, & M. Stede (2011): "Lexicon-basedmethods for sentiment analysis." *Computational Linguistics* **37(2)**: pp. 267–307.

Tey, C. M., P. Gupta, & D. Gao (2013): "I can be You: Questioning the use of Keystroke Dynamics as Biometrics." *20th Annual Network and Distributed System Security Symposium - NDSS '13* pp. 1 – 16.

Thelwall, M., K. Buckley, & G. Paltoglou (2011): "Sentiment in Twitter events." *Journal of the American Society for Information Science and Technology* .

Tumasjan, A., T. O. Sprenger, P. G. Sandner, & I. M. Welpe (2010): "Predicting elections with Twitter: What 140 characters reveal about political sentiment." In "ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media," pp. 178–185.

Ur, B., P. Kelly, S. Komanduri, J. Lee, M. Maass, M. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, & L. Cranor (2012): "How does your password measure up? the effect of strength meters on password creation." *Proc. Security '12, USENIX Association* .

Ur, B., S. Kom, R. Shay, S. Matsumoto, L. Bauer, N. Christin, L. F. Cranor, P. G. Kelley, M. L. Mazurek, & T. Vidas (2013): "Poster: The art of password creation." *Carnegie Mellon University* .

Ur, B., F. Noma, J. Bees, S. M. Segreti, R. Shay, L. Bauer, N. Christin, & L. F. Cranor (2019): ""i added '!' at the end to make it secure": Observing password creation in the lab." *SOUPS 2015 - Proceedings of the 11th Symposium on Usable Privacy and Security* .

Veras, R., C. Collins, & J. Thorpe (2014): "On the Semantic Patterns of Passwords and their Security Impact." *Internet Society* .

Wang, C., S. T. Jan, H. Hu, D. Bossart, & G. Wang (2018): "The next domino to fall: Empirical analysis of user passwords across online services." In "Proceedings of the Eighth ACM Conference on Data and Application

Security and Privacy," CODASPY '18, pp. 196–203. New York, NY, USA: ACM.

Wash, R., E. Rader, R. Berman, & Z. Wellmer (2016): "Understanding Password Choices: How Frequently Entered Passwords Are Re-used across Websites." *Twelfth Symposium on Usable Privacy and Security* .

Weber, J. E., D. Guster, P. Safonov, & M. B. Schmidt (2008): "Weak password security: An empirical study." *Information Security Journal* **17**.

Weir, M., S. Aggarwal, B. De Medeiros, & B. Glodek (2009): "Password cracking using probabilistic context-free grammars." In "Proceedings - IEEE Symposium on Security and Privacy," pp. 391–405.

Williams, R. (2016): "Understanding and interpreting generalized ordered logit models." *The Journal of Mathematical Sociology* **40(1)**: pp. 7–20.

Xu, L., C. Ge, W. Qiu, Z. Huang, Z. Gong, J. Guo, & H. Lian (2017): "Password guessing based on LSTM recurrent neural networks." In "Proceedings - 2017 IEEE International Conference on Computational Science and Engineering and IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, CSE and EUC 2017," volume 1, pp. 785–788. Institute of Electrical and Electronics Engineers Inc.

Yan, R. R. & H. Chen (2018): "Discovery and analysis of high frequency words in password sets." *Journal of Cryptologic Research* **5**.

Yi, J., T. Nasukawa, R. Bunescu, & W. Niblack (2003): "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques." In "Proceedings - IEEE International Conference on Data Mining, ICDM," pp. 427–434.

Yujian, L. & L. Bo (2007): "A normalized levenshtein distance metric." *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29(6)**: pp. 1091–1095.

Zeng, J. (2016): "China's date with big data: Will it strengthen or threaten authoritarian rule?" *International Affairs* **92(6)**: pp. 1443–1462.

Zeng, J., J. Duan, & C. Wu (2019): "Empirical study on lexical sentiment in passwords from chinese websites." *Computers & Security* **80**: pp. 200–210.

von Zezschwitz, E., A. De Luca, & H. Hussmann (2013): "Survival of the shortest: A retrospective analysis of influencing factors on password composition." In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler (editors), "Human-Computer Interaction – INTERACT 2013," pp. 460–467. Berlin, Heidelberg: Springer Berlin Heidelberg.

Zhang, M. L., Q. H. Zhang, W. F. Liu, X. X. Hu, & J. H. Wei (2019): "A Method of Password Attack Based on Structure Partition and String Reorganization." *Jisuanji Xuebao/Chinese Journal of Computers* **42(4)**.

Šolić, K., H. Očevčić, & D. Blažević (2015): "Survey on password quality and confidentiality." *Automatika – Journal for Control, Measurement, Electronics, Computing and Communications* **56**.

Appendix A

Accompanying tables and data

The appendix is organised as follows. First, it is shown the number the observations per TLD. Second, it is revealed the number of records per TLD in the cleaned data. Third, it is demonstrated how the Word Break Algorithm works. Third, it is given a comprehensive table indicating the estimated coefficients of Model family 1. Fourth, it is shown the estimated β coefficients of Model family 2. Last, a table indicating the significance of TLD for the Password-Password similarity is given.

A.1 Observations per TLD in the raw data

Table A.1: Counts of domains in the raw data

TLD	Count	TLD	Count	TLD	Count	TLD	Count	TLD	Count
com	844 200 121	eu	744 574	pe	78 772	xom	26 479	cojp	14 546
ru	226 594 848	nz	739 578	combr	77 941	ir	25 617	bw	14 479
de	63 271 746	ph	735 447	me	74 455	ke	25 234	ed	14 383
net	50 048 314	lv	711 160	nu	67 881	la	23 219	do	14 365
fr	44 986 923	co	675 422	si	67 857	comsg	22 872	comn	13 979
uk	26 644 014	mil	660 751	yu	61 702	mail	22 586	eg	13 870
it	24 675 740	sg	611 089	su	58 023	li	22 519	coop	13 809
pl	13 261 771	no	603 420	rrcom	56 818	coza	22 117	mt	13 383
cz	7 653 736	info	550 355	pk	56 097	mk	21 581	aol	13 302
cn	7 200 950	gov	514 006	comau	56 064	na	21 374	ug	13 250
edu	6 851 080	ie	472 657	lu	53 721	come	21 347	tc	13 249
jp	6 186 806	couk	458 289	uy	51 802	ms	21 302	uz	13 243
br	5 813 716	cl	454 581	az	50 243	name	20 600	qa	12 855
es	5 702 419	vn	443 848	kom	48 916	conz	20 311	py	12 854
ca	4 806 575	fi	433 321	ws	48 398	ec	20 279	com$_{a}buse$	12 709
ua	4 361 824	fm	430 922	rul	47 312	comhk	19 922	lru	12 470
au	3 918 009	il	385 353	sa	47 229	acuk	19 725	com1	12 329
org	3 596 060	kr	357 746	cpm	45 723	cu	19 555	bm	11 782
nl	3 342 572	tr	322 825	ge	43 322	comar	19 407	gt	11 440
in	3 200 151	lt	321 095	tn	42 708	ccom	19 239	armymil	11 289
hu	2 411 229	ry	284 297	ne	41 391	mu	19 019	pw	10 950
tw	1 982 277	con	271 889	ma	41 346	mobi	18 832	coil	10 881
mx	1 950 652	hr	270 866	comtw	40 910	commx	17 913	al	10 859
id	1 862 310	ee	253 925	ocm	39 087	to	17 790	zawq	10 727
at	1 565 892	msp	250 868	cat	37 901	ba	17 613	comq	10 605
sk	1 505 968	cc	234 956	comm	36 767	so	17 166	re	10 374
be	1 485 327	biz	230 524	tst	36 564	09	16 733	aero	10 326
	1 473 927	ro	209 540	commy	35 685	df	16 490	or	10 230
za	1 343 648	th	192 471	coom	33 887	am	16 319	as	10 222
dk	1 289 074	om	177 461	nf	32 732	cx	16 183	et	10 115
ch	1 255 912	cm	170 104	lk	30 235	comcn	16 070	tt	9 993
ar	1 223 460	coin	121 242	cy	29 829	jo	15 798	zm	9 835
se	1 190 279	tk	114 655	tu	29 130	ac	15 743	nc	9 268
bg	1 104 963	tv	111 239	cr	28 736	lb	15 649	np	9 141
my	1 095 997	coid	102 141	md	28 713	coml	15 444	ney	9 062
pt	1 089 181	rs	98 483	vom	28 264	yahoo	15 278	netau	8 790
by	1 027 310	is	94 731	cim	27 822	mz	15 205	live	8 668
gr	969 290	ae	89 970	ri	27 779	c0m	15 043	msn	8 531
hk	964 879	kz	88 595	ve	27 766	int	14 767	comvn	8 500
us	925 478	comph	86 533	zw	27 504	comw	14 641	Others	2 132 600

A.2 Number of observations per TLD in the cleaned data

Table A.2: Number of observation per country

TLD	Country Name	Population	Observations	Obs./Pop.
ad	Andorra	77,006	4,651	0.0604
ae	United Arab Emirates	9,630,959	88,689	0.0092
af	Afghanistan	37,172,384	1,860	0.0001
ag	Antigua and Barbuda	96,286	3,402	0.0353
al	Albania	2,866,376	10,793	0.0038
am	Armenia	2,951,776	16,147	0.0055
ao	Angola	30,809,762	6,456	0.0002
ar	Argentina	44,494,504	1,219,109	0.0274
as	American Samoa	55,465	10,128	0.1826
at	Austria	8,847,037	1,559,188	0.1762
au	Australia	24,992,368	3,883,094	0.1554
az	Azerbaijan	9,942,334	49,904	0.005
ba	Bosnia and Herzegovina	3,323,929	17,521	0.0053
bb	Barbados	286,641	2,102	0.0073
bd	Bangladesh	161,356,030	5,323	0.0
be	Belgium	11,422,068	1,476,400	0.1293
bf	Burkina Faso	19,751,536	1,808	0.0001
bg	Bulgaria	7,024,216	1,099,591	0.1565
bh	Bahrain	1,569,439	7,827	0.005
bi	Burundi	11,175,378	692	0.0001
bj	Benin	11,485,048	577	0.0001
bm	Bermuda	63,968	11,648	0.1821
bn	Brunei Darussalam	428,962	2,563	0.006
bo	Bolivia	11,353,142	6,813	0.0006
br	Brazil	209,469,330	5,762,338	0.0275
bs	Bahamas	385,640	2,208	0.0057
bt	Bhutan	754,394	2,200	0.0029

bw	Botswana	2,254,126	14,196	0.0063
by	Belarus	9,485,386	1,015,536	0.1071
bz	Belize	383,071	7,705	0.0201
ca	Canada	37,058,856	4,770,172	0.1287
cd	Democratic Republic of the Congo	84,068,090	3,192	0.0
cf	Central African Republic	4,666,377	2,791	0.0006
cg	Congo	5,244,363	541	0.0001
ch	Switzerland	8,516,543	1,247,758	0.1465
ci	Cote d'Ivoire	25,069,228	7,870	0.0003
cl	Chile	18,729,160	452,844	0.0242
cm	Cameroon	25,216,236	168,709	0.0067
cn	China	1,392,730,000	7,167,071	0.0051
co	Colombia	49,648,684	665,894	0.0134
cr	Costa Rica	4,999,441	28,594	0.0057
cu	Cuba	11,338,138	19,403	0.0017
cv	Cabo Verde	543,767	3,855	0.0071
cw	Curacao	159,849	146	0.0009
cy	Cyprus	1,189,265	29,566	0.0249
cz	Czechia	10,625,695	7,601,246	0.7154
de	Germany	82,927,920	63,080,765	0.7607
dj	Djibouti	958,920	5,485	0.0057
dk	Denmark	5,797,446	1,277,337	0.2203
dm	Dominica	71,625	722	0.0101
do	Dominican Republic	10,627,165	14,275	0.0013
dz	Algeria	42,228,428	6,746	0.0002
ec	Ecuador	17,084,356	20,008	0.0012
edu	usa -edu	327,167,420	6,767,896	0.0207
ee	Estonia	1,320,884	252,029	0.1908
eg	Egypt	98,423,590	13,589	0.0001
er	Eritrea	5,073,000	3,944	0.0008
es	Spain	46,723,748	5,673,029	0.1214
et	Ethiopia	109,224,560	10,029	0.0001
fi	Finland	5,518,050	430,791	0.0781

fj	Fiji	883,483	7,658	0.0087
fm	Micronesia	112,640	427,123	3.7919
fo	Faroe Islands	48,497	7,451	0.1536
fr	France	66,987,244	44,762,457	0.6682
ga	Gabon	2,119,275	4,254	0.002
gd	Grenada	111,454	2,247	0.0202
ge	Georgia	3,731,000	43,091	0.0115
gh	Ghana	29,767,108	7,809	0.0003
gl	Greenland	56,025	4,334	0.0774
gm	Gambia	2,280,102	906	0.0004
gn	Guinea	12,414,318	291	0.0
gov	usa -gov	327,167,420	502,171	0.0015
gq	Equatorial Guinea	1,308,974	841	0.0006
gr	Greece	10,727,668	957,578	0.0893
gt	Guatemala	17,247,808	11,352	0.0007
gw	Guinea-Bissau	1,874,309	67	0.0
gy	Guyana	779,004	521	0.0007
hk	China, Hong Kong	7,451,000	954,199	0.1281
hn	Honduras	9,587,522	3,389	0.0004
hr	Croatia	4,089,400	269,648	0.0659
ht	Haiti	11,123,176	879	0.0001
hu	Hungary	9,768,785	2,397,293	0.2454
id	Indonesia	267,663,440	1,837,314	0.0069
ie	Ireland	4,853,506	469,666	0.0968
il	Israel	8,883,800	382,045	0.043
in	India	1,352,617,300	3,121,811	0.0023
iq	Iraq	38,433,600	173	0.0
ir	Iran	81,800,270	25,188	0.0003
is	Iceland	353,574	94,425	0.2671
it	Italy	60,431,284	24,558,155	0.4064
jm	Jamaica	2,934,855	2,617	0.0009
jo	Jordan	9,956,011	15,568	0.0016
jp	Japan	126,529,100	6,170,003	0.0488
ke	Kenya	51,393,010	24,878	0.0005
kg	Kyrgyzstan	6,315,800	6,374	0.001
kh	Cambodia	16,249,798	4,429	0.0003

ki	Kiribati	115,847	1,005	0.0087
km	Comoros	832,322	455	0.0005
kr	Republic of Korea	51,635,256	356,601	0.0069
kw	Kuwait	4,137,309	5,039	0.0012
ky	Cayman Islands	64,174	7,013	0.1093
kz	Kazakhstan	18,276,500	87,326	0.0048
la	Lao People's Democratic Republic	7,061,507	23,039	0.0033
lb	Lebanon	6,848,925	15,270	0.0022
lc	Saint Lucia	181,889	860	0.0047
li	Liechtenstein	37,910	22,376	0.5902
lk	Sri Lanka	21,670,000	29,326	0.0014
lr	Liberia	4,818,977	804	0.0002
ls	Lesotho	2,108,132	2,541	0.0012
lt	Lithuania	2,789,533	318,736	0.1143
lu	Luxembourg	607,728	53,139	0.0874
lv	Latvia	1,926,542	707,157	0.3671
ly	Libya	6,678,567	1,671	0.0003
ma	Morocco	36,029,136	40,235	0.0011
mc	Monaco	38,682	3,786	0.0979
md	Republic of Moldova	3,545,883	28,490	0.008
me	Montenegro	622,345	73,483	0.1181
mg	Madagascar	26,262,368	8,025	0.0003
mil	usa - mil	327,167,420	634,487	0.0019
ml	Mali	19,077,690	4,019	0.0002
mm	Myanmar	53,708,396	3,270	0.0001
mn	Mongolia	3,170,208	6,776	0.0021
mo	China, Macao	631,636	6,970	0.011
mr	Mauritania	4,403,319	693	0.0002
ms	Montserrat	5,900	21,093	3.5751
mt	Malta	483,530	13,287	0.0275
mu	Mauritius	1,265,303	18,831	0.0149
mv	Maldives	515,696	2,090	0.0041
mw	Malawi	18,143,316	2,325	0.0001

mx	Mexico	126,190,780	1,930,205	0.0153
my	Malaysia	31,528,584	1,084,262	0.0344
mz	Mozambique	29,495,962	15,155	0.0005
na	Namibia	2,448,255	20,967	0.0086
ne	Niger	22,442,948	40,843	0.0018
nf	Norfolk Island	1,841	32,683	17.7529
ng	Nigeria	195,874,740	5,350	0.0
ni	Nicaragua	6,465,513	6,164	0.001
nl	Netherlands	17,231,016	3,304,896	0.1918
no	Norway	5,314,336	598,334	0.1126
np	Nepal	28,087,872	8,843	0.0003
nu	Niue	1,400	66,968	47.8343
nz	New Zealand	4,885,500	733,179	0.1501
om	Oman	4,829,483	176,262	0.0365
pa	Panama	4,176,873	5,473	0.0013
pe	Peru	31,989,256	78,359	0.0024
pf	French Polynesia	277,679	6,436	0.0232
pg	Papua New Guinea	8,606,316	7,503	0.0009
ph	Philippines	106,651,920	726,726	0.0068
pk	Pakistan	212,215,020	54,959	0.0003
pl	Poland	37,978,548	13,209,188	0.3478
pr	Puerto Rico	3,195,153	2,169	0.0007
ps	State of Palestine	4,569,087	2,572	0.0006
pt	Portugal	10,281,762	1,081,567	0.1052
py	Paraguay	6,956,071	12,746	0.0018
qa	Qatar	2,781,677	12,540	0.0045
ro	Romania	19,473,936	207,063	0.0106
rs	Serbia	6,982,084	98,187	0.0141
ru	Russian Federation	144,478,050	224,529,891	1.5541
rw	Rwanda	12,301,939	2,230	0.0002
sa	Saudi Arabia	33,699,948	46,490	0.0014
sb	Solomon Islands	652,858	1,468	0.0022
sc	Seychelles	96,762	3,212	0.0332
sd	Sudan	41,801,532	4,909	0.0001
se	Sweden	10,183,175	1,180,426	0.1159
sg	Singapore	5,638,676	604,945	0.1073

si	Slovenia	2,067,372	67,466	0.0326
sk	Slovakia	5,447,011	1,503,205	0.276
sl	Sierra Leone	7,650,154	671	0.0001
sm	San Marino	33,785	3,394	0.1005
sn	Senegal	15,854,360	8,173	0.0005
so	Somalia	15,008,154	17,150	0.0011
sr	Suriname	575,991	1,098	0.0019
ss	South Sudan	10,975,920	2,377	0.0002
st	Sao Tome and Principe	211,028	6,549	0.031
sv	El Salvador	6,420,744	6,943	0.0011
sy	Syrian Arab Republic	16,906,284	3,751	0.0002
sz	Swaziland	1,136,191	3,377	0.003
td	Chad	15,477,751	121	0.0
tg	Togo	7,889,094	762	0.0001
th	Thailand	69,428,530	191,318	0.0028
tj	Tajikistan	9,100,837	1,497	0.0002
tk	Tokelau	1,411	113,755	80.6201
tl	Timor-Leste	1,267,972	694	0.0005
tm	Turkmenistan	5,850,908	6,492	0.0011
tn	Tunisia	11,565,204	41,692	0.0036
to	Tonga	103,197	17,541	0.17
tr	Turkey	82,319,730	320,952	0.0039
tt	Trinidad and Tobago	1,389,858	9,797	0.007
tv	Tuvalu	11,508	109,543	9.5189
tz	United Republic of Tanzania	56,318,348	5,629	0.0001
ua	Ukraine	44,622,516	4,304,814	0.0965
ug	Uganda	42,723,140	13,129	0.0003
uk	United Kingdom	66,488,990	26,471,527	0.3981
us	United States of America	327,167,420	917,441	0.0028
uy	Uruguay	3,449,299	51,560	0.0149
uz	Uzbekistan	32,955,400	13,157	0.0004

vc	Saint Vincent and the Grenadines	110,210	1,161	0.0105
ve	Venezuela	28,870,196	27,623	0.001
vg	British Virgin Islands	29,802	1,577	0.0529
vn	Viet Nam	95,540,390	439,989	0.0046
vu	Vanuatu	292,680	4,068	0.0139
ws	Samoa	196,130	47,451	0.2419
ye	Yemen	28,498,688	2,010	0.0001
za	South Africa	57,779,624	1,321,816	0.0229
zm	Zambia	17,351,822	9,683	0.0006
zw	Zimbabwe	14,439,018	27,131	0.0019

Table A.3: Counts of the variables per country

TLD	**Country**	**Records**	**Providers**	**Users**	**Recurrent users**
ad	Andorra	4,651	667	3,913	574
ae	United Arab Emirates	88,689	9,741	81,484	4,913
af	Afghanistan	1,860	579	1,734	90
ag	Antigua and Barbuda	3,402	1,182	3,043	291
al	Albania	10,793	1,015	9,537	919
am	Armenia	16,147	2,251	13,622	1,648
ao	Angola	6,456	662	5,863	545
ar	Argentina	1,219,109	65,107	1,093,465	93,110
as	American Samoa	10,128	4,154	8,529	997
at	Austria	1,559,188	72,602	1,256,706	207,086
au	Australia	3,883,094	257,470	3,358,027	393,323
az	Azerbaijan	49,904	2,574	43,091	4,824
ba	Bosnia and Herzegovina	17,521	2,383	15,172	1,708
bb	Barbados	2,102	544	1,846	140
bd	Bangladesh	5,323	879	4,916	300
be	Belgium	1,476,400	89,462	1,177,433	181,029

Table A.3 continued from previous page

TLD	Country	Records	Providers	Users	Recurrent users
bf	Burkina Faso	1,808	229	1,593	181
bg	Bulgaria	1,099,591	5,847	898,694	136,055
bh	Bahrain	7,827	719	6,978	558
bi	Burundi	692	417	653	34
bj	Benin	577	164	533	38
bm	Bermuda	11,648	852	10,022	1,194
bn	Brunei Darussalam	2,563	459	2,336	188
bo	Bolivia	6,813	1,435	6,320	299
br	Brazil	5,762,338	287,442	5,039,652	500,455
bs	Bahamas	2,208	240	1,655	336
bt	Bhutan	2,200	330	1,956	156
bw	Botswana	14,196	1,104	12,498	1,505
by	Belarus	1,015,536	13,144	886,175	97,202
bz	Belize	7,705	2,891	6,649	701
ca	Canada	4,770,172	153,771	3,978,107	522,307
cd	Democratic Republic of the Congo	3,192	868	2,716	412
cf	Central African Republic	2,791	492	2,683	98
cg	Congo	541	247	455	73
ch	Switzerland	1,247,758	112,040	1,034,782	156,590
ci	Cote d'Ivoire	7,870	870	6,962	808
cl	Chile	452,844	48,184	409,086	31,070
cm	Cameroon	168,709	6,681	157,009	9,872
cn	China	7,167,071	107,228	6,378,743	650,001
co	Colombia	665,894	88,896	629,123	25,792
cr	Costa Rica	28,594	2,019	25,873	2,131
cu	Cuba	19,403	2,826	17,402	1,627
cv	Cabo Verde	3,855	624	3,428	375
cw	Curacao	146	80	132	14
cy	Cyprus	29,566	2,057	24,300	3,821
cz	Czechia	7,601,246	92,235	6,094,203	823,177
de	Germany	63,080,765	744,031	50,429,761	4,885,056

Table A.3 continued from previous page

TLD	Country	Records	Providers	Users	Recurrent users
dj	Djibouti	5,485	864	5,098	279
dk	Denmark	1,277,337	133,910	1,068,483	142,548
dm	Dominica	722	139	669	37
do	Dominican Republic	14,275	1,956	13,355	694
dz	Algeria	6,746	946	6,202	457
ec	Ecuador	20,008	3,918	19,010	808
edu	usa -edu	6,767,896	38,923	6,045,968	591,797
ee	Estonia	252,029	13,612	200,190	35,598
eg	Egypt	13,589	1,788	12,383	888
er	Eritrea	3,944	1,730	3,588	125
es	Spain	5,673,029	99,006	4,789,210	625,396
et	Ethiopia	10,029	1,360	9,621	328
fi	Finland	430,791	31,976	372,794	43,220
fj	Fiji	7,658	841	7,028	513
fm	Micronesia	427,123	3,275	325,184	63,306
fo	Faroe Islands	7,451	841	6,483	754
fr	France	44,762,457	169,911	34,347,575	6,684,037
ga	Gabon	4,254	628	4,017	209
gd	Grenada	2,247	494	1,771	404
ge	Georgia	43,091	3,610	37,694	4,107
gh	Ghana	7,809	1,556	7,100	493
gl	Greenland	4,334	475	3,812	412
gm	Gambia	906	195	800	81
gn	Guinea	291	72	253	31
gov	usa -gov	502,171	9,960	460,322	34,299
gq	Equatorial Guinea	841	133	796	20
gr	Greece	957,578	32,291	769,569	127,048
gt	Guatemala	11,352	2,396	10,546	547
gw	Guinea-Bissau	67	50	61	6
gy	Guyana	521	216	478	27

Table A.3 continued from previous page

TLD	Country	Records	Providers	Users	Recurrent users
hk	China, Hong Kong Special Administrative Region	954,199	19,400	794,467	113,076
hn	Honduras	3,389	627	3,176	173
hr	Croatia	269,648	11,996	230,527	27,806
ht	Haiti	879	423	834	29
hu	Hungary	2,397,293	52,658	1,765,078	290,968
id	Indonesia	1,837,314	14,101	1,660,814	129,743
ie	Ireland	469,666	32,586	369,897	59,993
il	Israel	382,045	24,342	334,229	37,409
in	India	3,121,811	56,095	2,773,469	264,510
iq	Iraq	173	73	152	11
ir	Iran	25,188	6,186	22,339	1,910
is	Iceland	94,425	7,374	83,698	8,107
it	Italy	24,558,155	331,975	18,632,106	3,751,850
jm	Jamaica	2,617	292	2,514	82
jo	Jordan	15,568	1,773	13,886	1,190
jp	Japan	6,170,003	102,210	4,948,527	725,173
ke	Kenya	24,878	4,616	22,771	1,788
kg	Kyrgyzstan	6,374	759	5,585	605
kh	Cambodia	4,429	935	4,186	199
ki	Kiribati	1,005	538	943	48
km	Comoros	455	177	358	83
kr	Republic of Korea	356,601	25,347	327,312	22,440
kw	Kuwait	5,039	629	4,581	307
ky	Cayman Islands	7,013	840	6,112	693
kz	Kazakhstan	87,326	7,577	79,967	5,831
la	Lao People's Democratic Republic	23,039	2,728	20,716	1,695
lb	Lebanon	15,270	996	13,457	1,396
lc	Saint Lucia	860	110	767	72
li	Liechtenstein	22,376	2,904	18,563	2,772

Table A.3 continued from previous page

TLD	Country	Records	Providers	Users	Recurrent users
lk	Sri Lanka	29,326	2,930	26,147	2,480
lr	Liberia	804	275	783	16
ls	Lesotho	2,541	376	2,260	248
lt	Lithuania	318,736	16,734	270,923	33,564
lu	Luxembourg	53,139	5,468	44,043	6,801
lv	Latvia	707,157	12,694	541,418	108,910
ly	Libya	1,671	779	1,501	90
ma	Morocco	40,235	5,396	36,006	3,510
mc	Monaco	3,786	633	3,285	351
md	Republic of Moldova	28,490	2,888	24,326	3,034
me	Montenegro	73,483	14,091	67,321	4,188
mg	Madagascar	8,025	618	6,271	708
mil	usa - mil	634,487	5,237	561,309	60,169
mk	The former Yugoslav Republic of Macedonia	21,244	2,802	18,542	1,907
ml	Mali	4,019	801	3,854	133
mm	Myanmar	3,270	626	2,859	297
mn	Mongolia	6,776	1,370	6,125	452
mo	China, Macao Special Administrative Region	6,970	788	6,340	464
mr	Mauritania	693	225	639	50
ms	Montserrat	21,093	1,392	19,003	1,650
mt	Malta	13,287	1,299	11,367	1,422
mu	Mauritius	18,831	1,488	15,621	2,396
mv	Maldives	2,090	539	1,899	152
mw	Malawi	2,325	238	2,111	184
mx	Mexico	1,930,205	53,212	1,751,148	141,072
my	Malaysia	1,084,262	25,191	1,007,430	55,787
mz	Mozambique	15,155	1,074	13,257	1,694
na	Namibia	20,967	1,441	18,014	2,407

Table A.3 continued from previous page

TLD	Country	Records	Providers	Users	Recurrent users
ne	Niger	40,843	7,425	38,025	2,040
nf	Norfolk Island	32,683	248	32,027	598
ng	Nigeria	5,350	870	4,978	302
ni	Nicaragua	6,164	988	5,782	299
nl	Netherlands	3,304,896	313,296	2,786,633	366,308
no	Norway	598,334	57,261	505,438	62,207
np	Nepal	8,843	1,707	8,135	566
nu	Niue	66,968	17,238	57,984	6,650
nz	New Zealand	733,179	57,906	637,017	68,404
om	Oman	176,262	7,019	166,450	8,268
pa	Panama	5,473	1,007	5,187	215
pe	Peru	78,359	7,720	73,597	3,101
pf	French Polynesia	6,436	494	5,655	621
pg	Papua New Guinea	7,503	768	7,081	353
ph	Philippines	726,726	9,086	644,197	63,668
pk	Pakistan	54,959	5,728	50,048	3,920
pl	Poland	13,209,188	130,334	10,462,959	1,672,662
pr	Puerto Rico	2,169	387	2,043	103
ps	State of Palestine	2,572	698	2,311	186
pt	Portugal	1,081,567	33,278	871,422	151,398
py	Paraguay	12,746	2,717	11,354	766
qa	Qatar	12,540	739	11,419	725
ro	Romania	207,063	44,054	178,638	18,234
rs	Serbia	98,187	7,799	84,874	9,227
ru	Russian Federation	224,529,891	539,789	178,008,405	25,516,318
rw	Rwanda	2,230	336	2,106	110
sa	Saudi Arabia	46,490	4,262	41,764	3,307
sb	Solomon Islands	1,468	162	1,355	100
sc	Seychelles	3,212	759	2,935	227
sd	Sudan	4,909	1,976	4,167	510
se	Sweden	1,180,426	120,711	985,454	135,461
sg	Singapore	604,945	19,354	509,808	68,586
si	Slovenia	67,466	10,465	60,020	5,818

Table A.3 continued from previous page

TLD	Country	Records	Providers	Users	Recurrent users
sk	Slovakia	1,503,205	41,402	1,102,864	163,834
sl	Sierra Leone	671	200	643	23
sm	San Marino	3,394	457	2,727	472
sn	Senegal	8,173	643	7,121	831
so	Somalia	17,150	753	14,620	2,375
sr	Suriname	1,098	351	1,061	33
ss	South Sudan	2,377	809	2,041	203
st	Sao Tome and Principe	6,549	1,822	5,566	640
sv	El Salvador	6,943	1,249	6,538	315
sy	Syrian Arab Republic	3,751	240	3,429	250
sz	Swaziland	3,377	392	2,984	349
td	Chad	121	69	110	10
tg	Togo	762	248	682	61
th	Thailand	191,318	9,659	177,012	11,469
tj	Tajikistan	1,497	352	1,367	99
tk	Tokelau	113,755	17,966	109,740	2,684
tl	Timor-Leste	694	224	661	31
tm	Turkmenistan	6,492	314	5,887	443
tn	Tunisia	41,692	3,093	34,695	5,210
to	Tonga	17,541	3,565	14,340	1,920
tr	Turkey	320,952	31,740	285,234	26,593
tt	Trinidad and Tobago	9,797	1,131	8,746	788
tv	Tuvalu	109,543	25,783	96,135	10,006
tz	United Republic of Tanzania	5,629	1,097	5,148	382
ua	Ukraine	4,304,814	80,274	3,884,463	316,062
ug	Uganda	13,129	1,425	11,383	1,356
uk	United Kingdom of Great Britain and Northern Ireland	26,471,527	1,025,452	22,014,131	3,095,395

Table A.3 continued from previous page

TLD	Country	Records	Providers	Users	Recurrent users
us	United States of America	917,441	77,800	823,206	71,914
uy	Uruguay	51,560	4,837	46,861	3,532
uz	Uzbekistan	13,157	1,423	11,650	1,139
vc	Saint Vincent and the Grenadines	1,161	535	1,036	104
ve	Venezuela	27,623	4,500	26,035	1,101
vg	British Virgin Islands	1,577	577	1,415	122
vn	Viet Nam	439,989	11,846	396,853	27,039
vu	Vanuatu	4,068	1,425	3,753	257
ws	Samoa	47,451	16,230	39,857	4,981
ye	Yemen	2,010	164	1,762	197
za	South Africa	1,321,816	149,042	1,156,517	134,003
zm	Zambia	9,683	839	8,701	854
zw	Zimbabwe	27,131	3,105	24,153	2,649

Table A.4: Gender identification per domain

TLD	Count	Females	Males	TLD	Count	Females	Males
ad	4664	304	528	lc	866	158	282
ae	89605	4028	17809	li	22422	1761	3492
af	1932	2	332	lk	29892	1283	2530
ag	3413	584	1045	lr	807	131	227
al	10824	174	489	ls	2597	310	620
am	16224	368	2397	lt	318971	20350	25673
ao	6489	716	2510	lu	53267	6681	12586
ar	1221242	219225	330021	lv	708108	67942	37923
as	10145	1404	2256	ly	1688	222	379
at	1560949	272831	387030	ma	41190	2437	8704
au	3899019	796724	1399135	mc	3793	304	569
az	49945	1843	6052	md	28518	3083	5690
ba	17555	2345	3166	me	73823	10793	16489
bb	2112	246	518	mg	8046	397	920

Table A.4 continued from previous page

TLD	Count	Females	Males	TLD	Count	Females	Males
bd	5415	53	687	mil	652742	104280	406145
be	1479829	273731	570852	mk	21479	2282	5355
bf	1817	304	469	ml	4040	401	953
bg	1101179	96428	123948	mm	3288	331	342
bh	7909	418	1869	mn	6806	162	67
bi	695	106	216	mo	6991	434	310
bj	582	85	103	mr	694	109	133
bm	11697	2033	4327	ms	21154	2732	4030
bn	2591	518	524	mt	13337	3020	4313
bo	6845	956	1366	mu	18943	3499	4588
br	5797285	923606	1446081	mv	2133	301	360
bs	2229	263	514	mw	2348	440	633
bt	2230	428	428	mx	1944274	292068	461145
bw	14455	2179	3486	my	1093057	11844	18146
by	1011910	224293	200986	mz	15186	1650	3734
bz	7738	1296	2382	na	21212	3856	5283
ca	4785007	1037298	1563100	ne	40864	2325	4224
cd	3207	97	386	nf	32685	3148	3209
cf	2799	162	261	ng	5397	990	1385
cg	541	73	90	ni	6184	771	1213
ch	1250063	192084	347641	nl	3328038	398824	785786
ci	7955	558	1351	no	599051	90258	192781
cl	453745	60400	98036	np	9093	97	1278
cm	169226	30350	41935	nu	67186	12688	18256
cn	7191598	180032	95788	nz	736037	151323	268397
co	667810	76344	91218	om	176980	4589	11199
cr	28645	3537	5282	pa	5482	693	1080
cu	19495	3020	4450	pe	78648	12422	12956
cv	3872	428	791	pf	6455	1493	1862
cw	146	16	15	pg	7671	1174	2086
cy	29663	748	1885	ph	729935	188592	211853
cz	7604766	858397	1044430	pk	55909	8657	8325
de	63120510	6852666	9097101	pl	13221354	1773725	1760347
dj	5501	494	937	pr	2205	424	734
dk	1278286	163628	335966	ps	2672	483	437
dm	723	116	255	pt	1084008	191546	329721

Table A.4 continued from previous page

TLD	Count	Females	Males	TLD	Count	Females	Males
do	14329	1909	3030	py	12784	1568	2590
dz	6848	410	1552	qa	12804	536	2992
ec	20087	2813	3648	ro	208004	36440	60537
edu	6801116	944432	2032367	rs	98371	21697	19841
ee	252132	16970	14216	ru	224690876	24674440	22943309
eg	13830	727	5318	rw	2241	499	541
er	3944	92	121	sa	47043	1801	11987
es	5685672	1030609	1334730	sb	1478	145	201
et	10076	186	381	sc	3220	622	1160
fi	431153	110802	168234	sd	4927	316	303
fj	7723	1254	1738	se	1182126	209610	410234
fm	427201	76606	101587	sg	607720	126733	148052
fo	7459	1033	1410	si	67551	14597	23080
fr	44842991	5298876	7200682	sk	1503802	222679	291474
ga	4276	756	1131	sl	673	90	125
gd	2247	95	170	sm	3402	1036	927
ge	43200	5509	6814	sn	8197	255	444
gh	7942	116	517	so	17155	140	320
gl	4335	606	1129	sr	1105	220	315
gm	912	152	174	ss	2381	96	128
gn	297	34	61	st	6572	1033	1785
gov	508733	122481	190525	sv	6982	1173	2050
gq	845	95	238	sy	3778	183	737
gr	966219	6827	13711	sz	3432	658	879
gt	11399	1376	2257	td	121	17	19
gw	67	3	5	tg	761	121	159
gy	526	62	137	th	192049	1409	935
hk	957549	34396	3041	tj	1501	287	334
hn	3410	514	660	tk	114017	18004	27854
hr	269984	45408	73367	tl	696	103	193
ht	888	98	177	tm	6501	2	6
hu	2399104	340205	468975	tn	42528	2668	9487
id	1857472	50192	169041	to	17595	2547	3687
ie	470852	104973	192444	tr	321588	44497	119222
il	383017	43769	83005	tt	9934	1695	3145
in	3184239	584999	1317766	tv	109830	17036	30238

Table A.4 continued from previous page

TLD	Count	Females	Males	TLD	Count	Females	Males
iq	177	6	3	tz	5699	1045	1502
ir	25447	907	4692	ua	4300844	196919	164030
is	94488	12680	16776	ug	13238	2576	3566
it	24614229	4025629	5818622	uk	26535484	5864979	9217659
jm	2651	549	1241	us	921769	173009	311331
jo	15757	1216	3838	uy	51623	6948	10528
jp	6176527	855602	425522	uz	13145	46	253
ke	25190	4652	5845	vc	1163	134	262
kg	6381	196	494	ve	27708	3844	6624
kh	4486	53	52	vg	1581	418	482
ki	1008	7	34	vn	442946	59865	79895
km	453	11	36	vu	4097	888	953
kr	357237	1032	2877	ws	47669	8534	14573
kw	5114	287	1372	ye	2018	315	661
ky	7040	1245	2446	za	1341143	279693	387767
kz	87245	8190	9146	zm	9818	1522	3174
la	23107			zw	27463	4729	6978
lb	15594	833	2305				

Table A.5: Descriptive statistics - Core data 1 - The similarity of a username and a password

TLD	Min	Mean	Max	SD	TLD	Min	Mean	Max	SD
ad	0,00	0,87	1,00	0,18	**lc**	0,00	0,87	1,00	0,17
ae	0,00	0,86	1,00	0,19	**li**	0,00	0,88	1,00	0,16
af	0,00	0,87	1,00	0,17	**lk**	0,00	0,84	1,00	0,21
ag	0,00	0,88	1,00	0,16	**lr**	0,00	0,85	1,00	0,26
al	0,00	0,88	1,00	0,16	**ls**	0,00	0,87	1,00	0,17
am	0,00	0,86	1,00	0,21	**lt**	0,00	0,85	1,00	0,19
ao	0,00	0,87	1,00	0,14	**lu**	0,00	0,88	1,00	0,15
ar	0,00	0,87	1,00	0,16	**lv**	0,00	0,85	1,00	0,18
as	0,00	0,87	1,00	0,17	**ly**	0,00	0,89	1,00	0,17
at	0,00	0,87	1,00	0,15	**ma**	0,00	0,87	1,00	0,20
au	0,00	0,87	1,00	0,15	**mc**	0,00	0,88	1,00	0,16
az	0,00	0,88	1,00	0,19	**md**	0,00	0,88	1,00	0,18
ba	0,00	0,86	1,00	0,17	**me**	0,00	0,90	1,00	0,15

Table A.5 continued from previous page

TLD	Min	Mean	Max	SD	TLD	Min	Mean	Max	SD
bb	0,00	0,85	1,00	0,19	**mg**	0,00	0,87	1,00	0,18
bd	0,00	0,85	1,00	0,21	**mk**	0,00	0,86	1,00	0,18
be	0,00	0,88	1,00	0,14	**ml**	0,00	0,90	1,00	0,13
bf	0,00	0,88	1,00	0,14	**mm**	0,00	0,87	1,00	0,20
bg	0,00	0,88	1,00	0,18	**mn**	0,00	0,88	1,00	0,18
bh	0,00	0,87	1,00	0,19	**mo**	0,00	0,89	1,00	0,14
bi	0,00	0,90	1,00	0,15	**mr**	0,00	0,88	1,00	0,18
bj	0,00	0,89	1,00	0,16	**ms**	0,00	0,91	1,00	0,14
bm	0,00	0,88	1,00	0,15	**mt**	0,00	0,87	1,00	0,16
bn	0,00	0,89	1,00	0,17	**mu**	0,00	0,87	1,00	0,17
bo	0,00	0,87	1,00	0,18	**mv**	0,00	0,86	1,00	0,19
br	0,00	0,89	1,00	0,16	**mw**	0,00	0,86	1,00	0,17
bs	0,00	0,89	1,00	0,16	**mx**	0,00	0,85	1,00	0,17
bt	0,00	0,86	1,00	0,19	**my**	0,00	0,86	1,00	0,17
bw	0,00	0,88	1,00	0,16	**mz**	0,00	0,87	1,00	0,16
by	0,00	0,87	1,00	0,22	**na**	0,00	0,87	1,00	0,18
bz	0,00	0,89	1,00	0,16	**ne**	0,00	0,88	1,00	0,17
ca	0,00	0,88	1,00	0,15	**nf**	0,00	0,94	1,00	0,09
cd	0,00	0,88	1,00	0,18	**ng**	0,00	0,86	1,00	0,17
cf	0,00	0,94	1,00	0,11	**ni**	0,00	0,87	1,00	0,18
cg	0,00	0,90	1,00	0,15	**nl**	0,00	0,87	1,00	0,15
ci	0,00	0,88	1,00	0,14	**no**	0,00	0,88	1,00	0,15
cl	0,00	0,87	1,00	0,16	**np**	0,00	0,83	1,00	0,21
cm	0,00	0,86	1,00	0,17	**nu**	0,00	0,88	1,00	0,15
cn	0,00	0,88	1,00	0,21	**nz**	0,00	0,87	1,00	0,15
co	0,00	0,88	1,00	0,16	**om**	0,00	0,86	1,00	0,17
cr	0,00	0,86	1,00	0,18	**pa**	0,00	0,88	1,00	0,17
cu	0,00	0,82	1,00	0,23	**pe**	0,00	0,88	1,00	0,16
cv	0,00	0,88	1,00	0,15	**pf**	0,00	0,87	1,00	0,16
cw	0,00	0,87	1,00	0,20	**pg**	0,00	0,87	1,00	0,17
cy	0,00	0,88	1,00	0,18	**ph**	0,00	0,86	1,00	0,16
cz	0,00	0,88	1,00	0,16	**pk**	0,00	0,86	1,00	0,19
de	0,00	0,88	1,00	0,13	**pl**	0,00	0,86	1,00	0,17
dj	0,00	0,92	1,00	0,14	**pr**	0,00	0,87	1,00	0,16
dk	0,00	0,88	1,00	0,16	**ps**	0,00	0,88	1,00	0,19
dm	0,00	0,87	1,00	0,16	**pt**	0,00	0,87	1,00	0,16

Table A.5 continued from previous page

TLD	Min	Mean	Max	SD	TLD	Min	Mean	Max	SD
do	0,00	0,88	1,00	0,16	**py**	0,00	0,88	1,00	0,18
dz	0,00	0,88	1,00	0,18	**qa**	0,00	0,88	1,00	0,17
ec	0,00	0,87	1,00	0,18	**ro**	0,00	0,86	1,00	0,16
ee	0,00	0,86	1,00	0,18	**rs**	0,00	0,87	1,00	0,16
eg	0,00	0,87	1,00	0,18	**ru**	0,00	0,87	1,00	0,20
er	0,00	0,89	1,00	0,18	**rw**	0,00	0,87	1,00	0,18
es	0,00	0,87	1,00	0,15	**sa**	0,00	0,88	1,00	0,17
et	0,00	0,87	1,00	0,17	**sb**	0,00	0,87	1,00	0,14
fi	0,00	0,87	1,00	0,14	**sc**	0,00	0,89	1,00	0,14
fj	0,00	0,87	1,00	0,17	**sd**	0,00	0,90	1,00	0,19
fm	0,00	0,87	1,00	0,18	**se**	0,00	0,88	1,00	0,14
fo	0,00	0,88	1,00	0,16	**sg**	0,00	0,88	1,00	0,17
fr	0,00	0,88	1,00	0,15	**si**	0,00	0,84	1,00	0,18
ga	0,00	0,90	1,00	0,15	**sk**	0,00	0,86	1,00	0,18
gd	0,00	0,88	1,00	0,19	**sl**	0,00	0,87	1,00	0,18
ge	0,00	0,85	1,00	0,20	**sm**	0,00	0,87	1,00	0,17
gh	0,00	0,87	1,00	0,19	**sn**	0,00	0,87	1,00	0,16
gl	0,00	0,89	1,00	0,15	**so**	0,00	0,90	1,00	0,20
gm	0,00	0,87	1,00	0,18	**sr**	0,00	0,87	1,00	0,16
gn	0,00	0,88	1,00	0,18	**ss**	0,00	0,87	1,00	0,21
gq	0,00	0,92	1,00	0,12	**st**	0,00	0,88	1,00	0,16
gr	0,00	0,89	1,00	0,16	**sv**	0,00	0,86	1,00	0,17
gt	0,00	0,86	1,00	0,19	**sy**	0,00	0,89	1,00	0,17
gw	0,56	0,91	1,00	0,11	**sz**	0,00	0,87	1,00	0,18
gy	0,00	0,88	1,00	0,18	**td**	0,36	0,91	1,00	0,13
hk	0,00	0,88	1,00	0,17	**tg**	0,00	0,88	1,00	0,18
hn	0,00	0,86	1,00	0,18	**th**	0,00	0,90	1,00	0,16
hr	0,00	0,86	1,00	0,17	**tj**	0,00	0,89	1,00	0,16
ht	0,00	0,88	1,00	0,16	**tk**	0,00	0,90	1,00	0,14
hu	0,00	0,88	1,00	0,16	**tl**	0,00	0,89	1,00	0,15
ch	0,00	0,88	1,00	0,14	**tm**	0,00	0,88	1,00	0,16
id	0,00	0,89	1,00	0,15	**tn**	0,00	0,87	1,00	0,20
ie	0,00	0,87	1,00	0,14	**to**	0,00	0,88	1,00	0,17
il	0,00	0,88	1,00	0,19	**tr**	0,00	0,90	1,00	0,15
in	0,00	0,86	1,00	0,16	**tt**	0,00	0,87	1,00	0,17
iq	0,31	0,89	1,00	0,15	**tv**	0,00	0,91	1,00	0,14

Table A.5 continued from previous page

TLD	Min	Mean	Max	SD	TLD	Min	Mean	Max	SD
ir	0,00	0,84	1,00	0,26	**tz**	0,00	0,86	1,00	0,18
is	0,00	0,84	1,00	0,22	**ua**	0,00	0,87	1,00	0,22
it	0,00	0,86	1,00	0,16	**ug**	0,00	0,85	1,00	0,19
jm	0,00	0,86	1,00	0,16	**uk**	0,00	0,87	1,00	0,15
jo	0,00	0,86	1,00	0,20	**us**	0,00	0,89	1,00	0,14
jp	0,00	0,85	1,00	0,21	**uy**	0,00	0,87	1,00	0,18
ke	0,00	0,86	1,00	0,19	**uz**	0,00	0,93	1,00	0,15
kg	0,00	0,90	1,00	0,17	**vc**	0,00	0,91	1,00	0,13
kh	0,00	0,89	1,00	0,20	**ve**	0,00	0,90	1,00	0,15
ki	0,00	0,91	1,00	0,17	**vg**	0,00	0,90	1,00	0,15
km	0,00	0,89	1,00	0,16	**vn**	0,00	0,88	1,00	0,17
kr	0,00	0,88	1,00	0,18	**vu**	0,00	0,89	1,00	0,16
kw	0,00	0,85	1,00	0,21	**ws**	0,00	0,88	1,00	0,17
ky	0,00	0,88	1,00	0,16	**ye**	0,00	0,90	1,00	0,16
kz	0,00	0,92	1,00	0,16	**za**	0,00	0,86	1,00	0,19
la	0,00	0,87	1,00	0,19	**zm**	0,00	0,87	1,00	0,17
lb	0,00	0,87	1,00	0,18	**zw**	0,00	0,86	1,00	0,18

Table A.6: Descriptive statistics - Core data 1 - Length of a password

TLD	Min	Mean	Max	SD	TLD	Min	Mean	Max	SD
ad	0,00	8,00	23,00	2,24	**lc**	0,00	8,30	21,00	2,52
ae	0,00	8,20	30,00	2,25	**li**	0,00	8,30	27,00	2,47
af	0,00	8,50	23,00	2,72	**lk**	0,00	8,10	30,00	2,50
ag	0,00	8,30	25,00	2,53	**lr**	0,00	8,20	19,00	2,38
al	0,00	8,50	30,00	2,31	**ls**	0,00	8,00	22,00	2,20
am	0,00	8,20	29,00	2,41	**lt**	0,00	8,20	30,00	2,04
ao	0,00	8,30	24,00	2,03	**lu**	0,00	8,10	29,00	2,16
ar	0,00	8,50	30,00	2,41	**lv**	0,00	8,40	30,00	2,19
as	0,00	7,80	24,00	2,20	**ly**	0,00	8,30	27,00	2,50
at	0,00	8,30	30,00	2,23	**ma**	0,00	8,20	30,00	2,36
au	0,00	8,00	30,00	2,08	**mc**	0,00	8,00	20,00	2,21
az	0,00	8,30	29,00	2,24	**md**	0,00	8,10	30,00	2,20
ba	0,00	8,00	28,00	2,21	**me**	0,00	8,90	30,00	2,04
bb	0,00	7,90	25,00	2,38	**mg**	0,00	7,70	26,00	2,35
bd	0,00	8,30	24,00	2,65	**mk**	0,00	8,20	24,00	2,34
be	0,00	8,20	30,00	2,01	**ml**	0,00	9,30	22,00	2,07
bf	0,00	7,90	19,00	1,96	**mm**	0,00	8,00	26,00	2,93
bg	0,00	8,00	30,00	2,14	**mn**	0,00	8,70	27,00	2,63
bh	0,00	7,90	26,00	2,17	**mo**	0,00	9,60	30,00	2,57
bi	0,00	8,00	20,00	2,29	**mr**	0,00	7,70	15,00	2,20
bj	0,00	7,80	17,00	2,03	**ms**	0,00	8,70	30,00	2,13
bm	0,00	8,10	26,00	2,28	**mt**	0,00	8,20	30,00	2,38
bn	0,00	8,00	18,00	2,37	**mu**	0,00	8,10	28,00	2,23
bo	0,00	8,80	29,00	2,88	**mv**	0,00	8,60	20,00	2,70
br	0,00	8,10	30,00	2,41	**mw**	0,00	8,40	21,00	2,47
bs	0,00	8,60	30,00	3,20	**mx**	0,00	8,50	30,00	2,18
bt	0,00	8,40	24,00	2,53	**my**	0,00	7,70	30,00	1,76
bw	0,00	8,30	24,00	2,36	**mz**	0,00	8,10	22,00	1,86
by	0,00	8,70	30,00	3,11	**na**	0,00	8,50	27,00	3,16
bz	0,00	8,10	29,00	2,16	**ne**	0,00	8,10	30,00	2,20
ca	0,00	8,30	30,00	2,31	**nf**	0,00	8,00	25,00	0,65
cd	0,00	7,90	17,00	2,22	**ng**	0,00	8,70	23,00	2,56
cf	0,00	9,40	26,00	1,59	**ni**	0,00	8,60	28,00	2,73
cg	0,00	7,70	18,00	1,92	**nl**	0,00	8,30	30,00	2,09
ci	0,00	8,00	30,00	1,95	**no**	0,00	8,00	30,00	2,09

Table A.6 continued from previous page

TLD	Min	Mean	Max	SD	TLD	Min	Mean	Max	SD
cl	0,00	8,70	30,00	2,42	**np**	0,00	8,60	22,00	2,54
cm	0,00	8,10	30,00	1,90	**nu**	0,00	8,00	30,00	1,98
cn	0,00	8,20	30,00	2,20	**nz**	0,00	8,10	30,00	2,18
co	0,00	8,20	30,00	2,65	**om**	0,00	8,10	30,00	1,91
cr	0,00	8,40	26,00	2,44	**pa**	0,00	8,80	26,00	2,65
cu	0,00	8,60	30,00	2,65	**pe**	0,00	8,50	28,00	2,59
cv	0,00	8,30	22,00	2,11	**pf**	0,00	7,90	23,00	2,37
cw	0,00	8,10	15,00	2,31	**pg**	0,00	8,30	22,00	2,60
cy	0,00	7,90	30,00	2,23	**ph**	0,00	8,10	30,00	2,17
cz	0,00	8,00	30,00	1,87	**pk**	0,00	8,40	30,00	2,45
de	0,00	9,40	30,00	3,15	**pl**	0,00	8,20	30,00	1,87
dj	0,00	7,80	22,00	1,55	**pr**	0,00	8,50	21,00	2,44
dk	0,00	8,20	30,00	2,14	**ps**	0,00	8,70	24,00	2,62
dm	1,00	8,00	23,00	2,12	**pt**	0,00	8,10	30,00	2,06
do	0,00	8,40	30,00	3,01	**py**	0,00	8,40	30,00	2,71
dz	0,00	8,30	23,00	2,33	**qa**	0,00	8,30	22,00	2,37
ec	0,00	8,90	27,00	2,76	**ro**	0,00	8,40	30,00	2,39
ee	0,00	8,00	30,00	2,09	**rs**	0,00	8,30	30,00	1,90
eg	0,00	8,30	27,00	2,50	**ru**	0,00	8,60	30,00	2,89
er	0,00	7,80	30,00	2,39	**rw**	0,00	8,50	28,00	2,36
es	0,00	8,40	30,00	2,05	**sa**	0,00	8,20	26,00	2,28
et	0,00	7,80	25,00	2,09	**sb**	0,00	8,20	23,00	3,08
fi	0,00	8,10	30,00	1,97	**sc**	0,00	9,00	23,00	2,47
fj	0,00	8,30	23,00	2,44	**sd**	0,00	7,40	25,00	2,49
fm	0,00	8,10	30,00	2,37	**se**	0,00	8,20	30,00	2,08
fo	0,00	8,00	26,00	1,95	**sg**	0,00	8,10	30,00	2,17
fr	0,00	8,10	30,00	1,92	**si**	0,00	7,90	29,00	2,02
ga	0,00	9,90	29,00	2,34	**sk**	0,00	7,90	30,00	1,87
gd	0,00	8,10	21,00	2,30	**sl**	0,00	7,70	16,00	1,98
ge	0,00	8,10	30,00	2,16	**sm**	0,00	8,30	26,00	2,32
gh	0,00	8,00	30,00	2,70	**sn**	0,00	7,90	22,00	2,03
gl	0,00	8,10	24,00	1,98	**so**	0,00	8,50	28,00	2,13
gm	1,00	8,30	18,00	2,23	**sr**	0,00	7,90	19,00	2,18
gn	0,00	8,10	17,00	2,19	**ss**	0,00	6,90	27,00	2,85
gq	0,00	9,60	25,00	1,63	**st**	0,00	8,00	30,00	2,11
gr	0,00	8,00	30,00	1,95	**sv**	0,00	8,90	30,00	2,83

Table A.6 continued from previous page

TLD	Min	Mean	Max	SD	TLD	Min	Mean	Max	SD
gt	0,00	8,70	29,00	2,82	**sy**	0,00	7,90	21,00	2,68
gw	0,00	7,70	15,00	2,75	**sz**	0,00	8,20	21,00	2,31
gy	1,00	8,10	17,00	2,44	**td**	3,00	7,90	18,00	2,39
hk	0,00	8,00	30,00	2,29	**tg**	0,00	8,00	21,00	2,14
hn	0,00	8,60	23,00	2,74	**th**	0,00	8,30	30,00	2,04
hr	0,00	8,10	30,00	2,03	**tj**	0,00	8,00	20,00	2,08
ht	0,00	8,80	18,00	2,41	**tk**	0,00	9,50	30,00	1,95
hu	0,00	7,90	30,00	1,59	**tl**	1,00	8,80	30,00	3,76
ch	0,00	8,10	30,00	2,10	**tm**	0,00	8,80	27,00	2,66
id	0,00	8,20	30,00	2,41	**tn**	0,00	7,70	30,00	2,40
ie	0,00	8,30	30,00	2,19	**to**	0,00	8,00	30,00	1,99
il	0,00	7,70	30,00	2,21	**tr**	0,00	8,20	30,00	2,23
in	0,00	8,50	30,00	2,42	**tt**	0,00	8,20	30,00	2,45
iq	0,00	9,00	20,00	3,39	**tv**	0,00	8,70	30,00	2,04
ir	0,00	8,40	30,00	2,59	**tz**	0,00	8,40	19,00	2,45
is	0,00	7,30	28,00	2,35	**ua**	0,00	8,60	30,00	3,04
it	0,00	8,10	30,00	2,08	**ug**	0,00	7,70	27,00	2,34
jm	0,00	8,50	23,00	2,53	**uk**	0,00	8,40	30,00	2,06
jo	0,00	8,20	24,00	2,30	**us**	0,00	8,00	30,00	2,17
jp	0,00	8,10	30,00	1,91	**uy**	0,00	8,30	25,00	2,34
ke	0,00	8,30	29,00	2,46	**uz**	0,00	9,00	28,00	2,57
kg	0,00	8,00	25,00	2,32	**vc**	0,00	8,20	23,00	2,19
kh	0,00	8,50	28,00	2,63	**ve**	0,00	8,50	27,00	2,49
ki	0,00	7,70	23,00	3,23	**vg**	0,00	7,70	28,00	2,13
km	0,00	8,20	19,00	2,59	**vn**	0,00	8,40	30,00	2,36
kr	0,00	7,20	30,00	2,14	**vu**	0,00	8,60	30,00	2,30
kw	0,00	8,00	24,00	2,16	**ws**	0,00	8,20	30,00	2,36
ky	0,00	8,70	25,00	2,34	**ye**	0,00	8,10	24,00	2,38
kz	0,00	8,90	30,00	2,79	**za**	0,00	7,90	30,00	2,16
la	0,00	8,10	28,00	1,98	**zm**	0,00	8,20	20,00	2,40
lb	0,00	8,30	24,00	2,40	**zw**	0,00	8,30	24,00	2,35

Table A.7: Descriptive statistics - Core data 1 - The Effort

TLD	Cat 1	Cat 2	Cat 3	Cat 4	TLD	Cat 1	Cat 2	Cat 3	Cat 4
ad	63,9%	28,8%	6,7%	0,5%	**lc**	54,2%	40,8%	4,4%	0,5%
ae	49,0%	42,7%	7,6%	0,6%	**li**	45,5%	45,9%	7,8%	0,6%
af	55,1%	28,1%	13,3%	2,5%	**lk**	59,1%	34,8%	5,3%	0,7%
ag	45,3%	45,6%	8,1%	0,7%	**lr**	61,8%	33,7%	4,1%	0,3%
al	63,1%	26,9%	8,9%	1,0%	**ls**	65,4%	26,9%	6,1%	1,4%
am	52,8%	39,4%	7,2%	0,4%	**lt**	54,2%	41,7%	3,8%	0,2%
ao	75,6%	19,1%	4,7%	0,4%	**lu**	53,9%	37,3%	7,9%	0,7%
ar	52,3%	41,7%	5,5%	0,5%	**lv**	48,6%	46,4%	4,4%	0,3%
as	49,0%	43,0%	7,4%	0,5%	**ly**	56,7%	34,9%	7,5%	0,8%
at	40,8%	53,3%	5,5%	0,3%	**ma**	69,0%	25,7%	4,6%	0,5%
au	32,1%	60,0%	7,3%	0,5%	**mc**	62,0%	33,4%	4,2%	0,3%
az	57,6%	35,5%	6,3%	0,4%	**md**	58,7%	35,1%	5,5%	0,5%
ba	61,6%	32,7%	5,1%	0,5%	**me**	31,0%	57,9%	10,2%	0,8%
bb	60,8%	33,1%	5,5%	0,5%	**mg**	79,1%	17,6%	2,8%	0,3%
bd	54,5%	37,7%	6,8%	0,9%	**mk**	54,3%	36,8%	7,6%	1,2%
be	60,0%	33,5%	5,8%	0,6%	**ml**	35,4%	52,7%	11,0%	0,8%
bf	79,0%	18,8%	1,8%	0,3%	**mm**	53,9%	40,0%	5,5%	0,2%
bg	50,9%	43,4%	5,1%	0,4%	**mn**	45,2%	36,1%	16,0%	2,5%
bh	56,3%	38,1%	4,5%	0,7%	**mo**	23,4%	31,3%	40,7%	4,3%
bi	53,7%	34,9%	10,7%	0,3%	**mr**	74,2%	22,0%	3,0%	0,3%
bj	74,2%	21,1%	3,2%	0,4%	**ms**	27,3%	42,4%	29,5%	0,7%
bm	48,4%	44,1%	6,8%	0,6%	**mt**	53,6%	38,6%	6,7%	1,0%
bn	50,9%	41,2%	7,1%	0,6%	**mu**	62,2%	32,0%	5,0%	0,6%
bo	59,4%	32,4%	7,2%	0,9%	**mv**	60,3%	30,3%	8,5%	0,7%
br	49,4%	43,0%	6,7%	0,8%	**mw**	59,4%	30,3%	9,2%	0,8%
bs	37,7%	35,9%	23,6%	1,7%	**mx**	38,6%	53,2%	7,4%	0,7%
bt	56,0%	38,4%	5,0%	0,3%	**my**	27,6%	69,5%	2,6%	0,1%
bw	63,7%	24,8%	10,4%	1,0%	**mz**	78,6%	17,3%	3,6%	0,3%
by	47,4%	43,1%	8,0%	0,7%	**na**	56,2%	35,3%	7,6%	0,8%
bz	35,5%	56,1%	7,3%	0,9%	**ne**	35,5%	45,2%	18,3%	0,6%
ca	39,9%	51,1%	8,2%	0,6%	**nf**	84,6%	13,7%	1,6%	0,0%
cd	68,5%	27,2%	3,7%	0,3%	**ng**	63,5%	28,5%	7,1%	0,8%
cf	12,9%	45,4%	41,1%	0,5%	**ni**	53,0%	39,1%	7,2%	0,7%
cg	65,1%	32,7%	1,5%	0,4%	**nl**	43,9%	44,0%	10,8%	1,1%
ci	82,2%	14,0%	3,2%	0,5%	**no**	36,3%	52,2%	10,8%	0,6%

Table A.7 continued from previous page

TLD	Cat 1	Cat 2	Cat 3	Cat 4	TLD	Cat 1	Cat 2	Cat 3	Cat 4
cl	42,7%	46,8%	9,4%	1,0%	**np**	55,4%	37,1%	6,6%	0,7%
cm	33,1%	63,5%	3,1%	0,1%	**nu**	47,1%	42,6%	8,9%	1,3%
cn	67,0%	29,4%	3,4%	0,1%	**nz**	33,6%	59,0%	6,8%	0,4%
co	35,1%	52,0%	9,7%	0,7%	**om**	32,5%	63,9%	3,4%	0,2%
cr	42,1%	47,0%	10,1%	0,6%	**pa**	45,4%	40,5%	12,7%	1,4%
cu	49,0%	42,0%	7,9%	0,7%	**pe**	63,5%	30,6%	5,4%	0,5%
cv	71,4%	20,6%	7,3%	0,7%	**pf**	60,6%	32,4%	6,1%	0,6%
cw	42,4%	54,5%	2,3%	0,0%	**pg**	54,8%	36,6%	7,7%	0,6%
cy	62,9%	31,8%	4,7%	0,5%	**ph**	53,6%	40,7%	5,0%	0,5%
cz	63,7%	31,2%	4,8%	0,3%	**pk**	57,0%	35,7%	6,5%	0,7%
de	44,6%	47,8%	7,3%	0,3%	**pl**	53,4%	40,6%	5,6%	0,3%
dj	28,2%	42,6%	29,0%	0,1%	**pr**	45,0%	46,5%	7,6%	0,9%
dk	44,6%	46,3%	8,6%	0,4%	**ps**	50,4%	37,1%	11,0%	1,2%
dm	46,8%	46,0%	6,9%	0,3%	**pt**	67,2%	27,5%	4,8%	0,5%
do	46,1%	46,3%	7,0%	0,5%	**py**	56,6%	37,3%	5,4%	0,5%
dz	68,4%	26,7%	4,0%	0,6%	**qa**	51,3%	38,8%	8,4%	1,4%
ec	52,2%	40,7%	6,2%	0,7%	**ro**	62,6%	28,7%	8,0%	0,6%
ee	48,2%	38,5%	12,8%	0,4%	**rs**	65,8%	29,3%	4,4%	0,4%
eg	60,7%	30,4%	7,5%	1,1%	**ru**	46,6%	46,2%	6,1%	0,3%
er	79,2%	18,8%	1,3%	0,0%	**rw**	62,6%	31,7%	4,5%	0,5%
es	61,1%	32,8%	5,4%	0,5%	**sa**	58,4%	34,5%	6,3%	0,6%
et	37,5%	58,0%	3,9%	0,3%	**sb**	30,8%	64,0%	4,9%	0,2%
fi	42,6%	45,0%	11,5%	0,7%	**sc**	35,3%	35,9%	27,3%	1,4%
fj	49,1%	40,8%	8,9%	0,9%	**sd**	75,3%	22,0%	2,3%	0,4%
fm	54,0%	40,0%	4,8%	0,5%	**se**	46,1%	43,9%	9,2%	0,7%
fo	35,6%	55,6%	8,2%	0,5%	**sg**	48,7%	45,5%	5,2%	0,4%
fr	65,0%	30,3%	4,1%	0,4%	**si**	48,0%	47,4%	4,2%	0,3%
ga	16,3%	42,7%	35,5%	5,4%	**sk**	62,2%	32,3%	5,2%	0,3%
gd	58,9%	35,8%	4,6%	0,3%	**sl**	61,7%	33,6%	4,3%	0,2%
ge	56,6%	39,2%	3,8%	0,3%	**sm**	64,3%	31,0%	4,3%	0,3%
gh	56,7%	35,4%	6,7%	1,1%	**sn**	81,2%	15,6%	2,8%	0,2%
gl	47,5%	43,9%	7,7%	0,7%	**so**	76,0%	19,6%	3,9%	0,3%
gm	62,6%	32,8%	3,9%	0,7%	**sr**	55,6%	37,9%	5,3%	0,7%
gn	71,2%	22,6%	4,3%	1,6%	**ss**	72,0%	26,6%	1,2%	0,1%
gq	13,9%	60,8%	24,3%	0,9%	**st**	39,6%	51,2%	8,3%	0,7%
gr	59,2%	36,8%	3,7%	0,3%	**sv**	47,3%	42,2%	9,1%	1,2%

Table A.7 continued from previous page

TLD	Cat 1	Cat 2	Cat 3	Cat 4	TLD	Cat 1	Cat 2	Cat 3	Cat 4
gt	54,4%	37,2%	7,5%	0,8%	**sy**	51,5%	38,2%	8,6%	0,8%
gw	41,0%	57,4%	0,0%	0,0%	**sz**	63,9%	28,4%	6,9%	0,6%
gy	52,6%	40,7%	6,2%	0,4%	**td**	70,9%	27,3%	1,8%	0,0%
hk	57,7%	36,6%	5,0%	0,4%	**tg**	71,4%	23,9%	4,4%	0,1%
hn	51,5%	40,1%	7,4%	0,8%	**th**	66,1%	23,6%	8,9%	1,2%
hr	42,8%	49,9%	6,7%	0,5%	**tj**	57,8%	32,0%	9,1%	0,5%
ht	40,7%	52,1%	6,5%	0,5%	**tk**	5,8%	87,3%	6,7%	0,2%
hu	65,6%	27,6%	6,3%	0,4%	**tl**	40,4%	52,2%	6,5%	0,9%
ch	49,3%	41,6%	8,3%	0,7%	**tm**	39,5%	54,0%	6,2%	0,3%
id	57,3%	32,0%	9,2%	1,2%	**tn**	69,6%	24,9%	4,9%	0,4%
ie	43,2%	48,7%	7,3%	0,7%	**to**	47,2%	45,7%	6,5%	0,3%
il	45,8%	47,6%	5,9%	0,5%	**tr**	67,6%	27,0%	5,0%	0,3%
in	58,9%	33,1%	7,0%	0,8%	**tt**	42,6%	47,8%	8,3%	1,1%
iq	43,2%	38,7%	16,1%	1,3%	**tv**	26,6%	42,9%	30,1%	0,4%
ir	60,9%	34,5%	3,9%	0,5%	**tz**	65,3%	25,1%	8,3%	1,0%
is	45,7%	48,6%	5,0%	0,4%	**ua**	48,4%	43,3%	6,7%	0,7%
it	67,2%	28,9%	3,6%	0,2%	**ug**	72,6%	21,8%	4,6%	0,6%
jm	48,2%	41,8%	9,0%	1,0%	**uk**	34,5%	58,2%	6,9%	0,3%
jo	53,9%	39,8%	5,5%	0,7%	**us**	33,2%	55,5%	10,3%	0,9%
jp	45,7%	50,7%	3,3%	0,2%	**uy**	42,2%	50,2%	6,8%	0,6%
ke	65,2%	25,8%	7,6%	1,0%	**uz**	39,7%	26,1%	33,8%	0,3%
kg	61,7%	30,6%	6,9%	0,5%	**vc**	52,7%	41,0%	5,5%	0,8%
kh	63,8%	29,3%	6,3%	0,5%	**ve**	55,2%	37,9%	6,2%	0,7%
ki	62,2%	31,4%	5,9%	0,2%	**vg**	44,4%	49,9%	5,1%	0,2%
km	63,3%	30,8%	5,3%	0,3%	**vn**	62,1%	30,8%	6,1%	0,7%
kr	46,1%	50,9%	1,8%	0,2%	**vu**	33,4%	52,6%	13,0%	0,8%
kw	53,2%	40,0%	6,2%	0,6%	**ws**	33,6%	53,3%	11,3%	0,5%
ky	36,6%	38,6%	24,2%	0,5%	**ye**	64,2%	29,8%	5,1%	0,7%
kz	44,2%	28,6%	26,3%	0,5%	**za**	51,2%	39,4%	8,1%	1,2%
la	22,7%	71,3%	5,6%	0,3%	**zm**	62,0%	27,7%	8,4%	1,0%
lb	59,6%	31,0%	8,3%	0,9%	**zw**	65,2%	27,1%	6,9%	0,6%

Table A.8 gives an overview of the macroeconomic variables. DEM stands for Democracy index, MOB for Mobile usage, NET for Internet usagem SEC for Cybersecurity index and LIT for Literacy rate.

Table A.8: Descriptive statistics - Macroeconomic variables

TLD	DEM	MOB	NET	SEC	LIT	TLD	DEM	MOB	NET	SEC	LIT
ad	107,3	91,6	0,12			**lc**	2,10	51,9	25,5	0,20	84,7
ae	2,52	208,5	94,8	0,81	93,2	**li**	5,82	64,5	78,2	0,19	95,1
af	2,48	59,1	13,5	0,18	43,0	**lk**	101,7	50,8	0,10		
ag	76,0	0,25	99,0			**lr**	124,7	98,1	0,54	99,0	53,5
al	5,86	94,2	71,8	0,63	98,1	**ls**	6,64	142,7	34,1	0,47	91,7
am	4,09	121,3	64,7	0,50	99,7	**lt**	5,07	8,0	0,21	48,3	91,3
ao	3,32	43,1	14,3	0,10	66,0	**lu**	6,02	29,0	0,05	76,6	
ar	6,84	132,1	74,3	0,41	99,0	**lv**	7,24	163,9	77,6	0,91	99,8
as	0,37					**ly**	8,88	132,2	97,4	0,89	99,0
at	8,49	123,5	87,9	0,83	99,0	**ma**	7,05	107,3	80,1	0,75	99,9
au	9,22	113,6	86,5	0,89	99,0	**mc**	1,94	21,8	0,21	86,1	60,7
az	3,15	103,9	79,0	0,65	99,8	**md**	3,79	124,2	61,8	0,43	73,8
ba	5,32	104,1	64,9	0,20	97,0	**me**	84,5	97,1	0,75	30,6	
bb	114,9	81,8	0,17	99,6		**mg**	6,33	88,0	76,1	0,66	73,8
bd	5,87	100,2	15,0	0,53	73,9	**mk**	6,27	180,7	71,3	0,64	98,8
be	8,05	99,7	87,7	0,81	99,0	**ml**	3,94	40,6	9,8	0,20	74,8
bf	3,59	97,9	16,0	0,40	41,2	**mm**	8,18	129,0	87,3	0,93	99,0
bg	6,84	118,9	63,4	0,72	98,4	**mn**	6,16	94,5	74,5	97,8	67,9
bh	3,49	133,3	95,9	0,59	97,5	**mo**	6,01	115,1	13,0	0,09	35,5
bi	4,01	56,5	2,7	0,09	68,4	**mr**	1,77	113,8	30,7	0,17	75,6
bj	6,17	82,4	20,0	0,49	42,4	**ms**	6,36	133,2	23,7	0,46	98,4
bm	98,4					**mt**	345,3	83,2	96,5	94,4	
bn	131,9	94,9	0,62	97,2		**mu**	3,86	103,7	20,8	0,11	53,5
bo	5,92	100,8	43,8	0,14	92,5	**mv**	8,28	140,2	81,0	0,48	94,5
br	7,12	98,8	67,5	0,58	93,2	**mw**	8,04	151,4	55,4	0,88	91,3
bs	99,0	85,0	0,15			**mx**	166,4	63,2	0,00	97,7	59,1
bt	4,68	93,3	48,1	0,18	66,6	**my**	5,84	39,0	13,8	0,28	62,1
bw	7,63	150,0	47,0	0,44	86,8	**mz**	6,93	95,2	63,9	0,63	95,4
by	3,34	122,9	74,4	0,58	99,8	**na**	6,19	134,5	80,1	0,89	94,9
bz	85,5	47,1	0,13	70,3		**ne**	4,90	47,7	10,0	0,16	60,7
ca	9,08	89,6	91,0	0,89	99,0	**nf**	6,23	112,7	51,0	0,13	91,5
cd	2,15	43,4	8,6	0,17	77,0	**ng**	3,38	10,2	0,09	30,6	93,5
cf	1,82	27,4	4,3	0,04	37,4	**ni**	3,47	88,2	42,0	0,65	62,0
cg	2,89	95,3	8,7	0,17	80,3	**nl**	5,73	115,1	27,9	0,13	82,6
ci	3,02	134,9	43,8	0,46	47,2	**no**	8,99	123,7	93,2	0,89	99,0
cl	7,67	134,4	82,3	0,47	96,4	**np**	9,80	107,2	96,4	0,89	99,0
cm	3,41	73,2	23,2	0,43	77,1	**nu**	4,24	139,4	34,0	0,26	67,9
cn	3,14	115,5	54,3	0,83	96,8	**nz**	9,26	134,9	90,8	0,79	
co	6,55	129,9	62,3	0,57	95,1	**om**	2,86	133,4	80,2	0,87	95,7
cr	8,04	169,9	71,4	0,22	97,9	**pa**	7,15	137,0	57,9	0,37	95,4
cu	3,52	47,4	57,1	0,48	99,8	**pe**	6,40	48,7	0,40	94,4	99,0
cv	7,94	112,2	57,2	0,05	86,8	**pf**	109,0	72,7	84,4	0,90	97,3
cw	114,5	68,1				**pg**	6,54	11,2	0,13	61,6	99,7
cy	7,29	138,9	80,7	0,65	98,7	**ph**	6,12	126,2	60,1	0,64	98,2
cz	8,19	119,1	78,7	0,57	99,0	**pk**	4,55	72,6	15,5	0,41	59,1
de	8,38	129,3	84,4	0,85	99,0	**pl**	7,05	134,7	76,0	0,82	98,7
dj	2,20	41,2	55,7	0,06		**pr**	109,6	70,6	92,4	0,31	

Table A.8 continued from previous page

TLD	DEM	MOB	NET	SEC	LIT	TLD	DEM	MOB	NET	SEC	LIT
dk	9,52	125,1	97,1	0,85	99,0	**ps**	90,0	65,2	0,31	97,2	
dm	105,8	69,6				**pt**	8,02	115,6	73,8	0,76	96,1
do	6,20	84,1	67,6	0,43	93,8	**py**	6,40	107,0	61,1	0,60	94,0
dz	3,44	111,7	47,7	0,26	81,4	**qa**	3,09	141,9	97,4	0,86	93,5
ec	5,77	92,3	57,3	0,37	92,8	**ro**	6,60	116,2	63,7	0,57	98,8
ee	8,18	129,0	87,3	0,93	99,0	**rs**	6,33	95,8	70,3	0,64	98,8
eg	7,68	145,4	88,1	0,91	99,9	**ru**	4,26	157,4	76,0	0,84	99,7
er	3,07	95,3	45,0	0,84	71,2	**rw**	3,25	78,9	21,8	0,70	73,2
es	2,31	1,3	0,02	76,6		**sa**	1,84	122,6	82,1	0,88	95,3
et	8,16	116,0	84,6	0,90	98,4	**sb**	73,8	11,9	0,06	76,6	93,8
fi	3,68	18,6	0,28	51,8		**sc**	184,3	58,8	0,26	95,9	
fj	9,19	129,5	87,5	0,86	99,0	**sd**	2,42	72,0	30,9	0,29	60,7
fm	3,62	50,0	0,19	99,1		**se**	9,50	126,8	95,5	0,81	99,0
fo	35,3	0,04				**sg**	5,89	148,8	84,4	0,90	97,3
fr	117,1	97,6				**si**	7,69	118,7	78,9	0,70	99,7
ga	7,77	108,4	80,5	0,92	99,0	**sk**	7,35	132,8	81,6	0,73	99,0
gd	3,29	138,3	62,0	0,32	84,7	**sl**	4,51	9,0	0,20	43,2	98,7
ge	59,1	0,14	98,6			**sm**	60,2	0,08	99,9		
gh	4,59	136,4	59,7	0,86	99,4	**sn**	5,27	104,5	46,0	0,31	77,9
gl	6,02	137,5	39,0	0,44	79,0	**so**	51,0	2,0	0,07	0,66	100,0
gm	110,6	69,5				**sr**	6,65	130,6	48,9	0,11	94,4
gn	3,38	139,5	19,8	0,28	50,8	**ss**	33,5	8,0	0,07	34,5	99,0
gq	2,79	96,8	18,0	0,19	32,0	**st**	77,1	29,9	0,06	92,8	99,0
gr	8,18	129,0	87,3	0,93	99,0	**sv**	6,47	146,9	33,8	0,12	89,0
gt	1,84	45,2	26,2	0,03	94,4	**sy**	2,31	101,1	34,3	0,24	80,8
gw	7,92	115,7	70,5	0,53	97,9	**sz**	2,90	47,0	0,13	88,4	
gy	6,05	118,7	65,0	0,25	81,3	**td**	1,52	45,1	6,5	0,10	22,3
hk	1,99	79,0	3,9	45,6		**tg**	3,45	77,9	12,4	63,7	
hn	6,05	37,3	0,13	85,6		**th**	6,55	180,2	52,9	0,80	93,8
hr	5,92	270,0	89,4	99,0		**tj**	2,51	22,0	0,26	99,8	
ht	5,76	79,2	31,7	0,04	87,2	**tk**	7,22	115,8	27,5	0,08	68,1
hu	6,81	105,6	67,1	0,84	99,1	**tl**	1,72	21,3	0,12	99,7	54,1
ch	4,00	57,5	12,3	0,05	61,7	**tm**	2,79	127,7	64,2	0,54	79,0
id	7,21	103,4	76,8	0,81	99,1	**tn**	104,6	41,2	0,21	99,4	86,7
ie	9,09	126,8	89,7	0,79	99,0	**to**	5,73	97,3	64,7	0,85	96,2
il	6,53	119,3	32,3	0,78	95,7	**tr**	7,16	141,9	77,3	0,19	98,7
in	8,79	103,2	84,1	0,78	99,0	**tt**	49,3	0,06			
iq	7,48	127,7	81,6	0,78	91,8	**tv**	5,64	77,2	25,0	0,64	77,9
ir	7,28	86,9	34,5	0,72	74,4	**tz**	6,30	127,8	58,9	0,66	100,0
is	4,00	95,0	49,4	0,26	85,6	**ua**	5,05	57,3	23,7	0,62	76,5
it	1,94	108,5	64,0	0,64	85,5	**ug**	8,16	118,4	94,6	0,93	99,0
jm	9,65	126,1	98,3	0,45	99,0	**uk**	8,18	129,0	87,3	0,93	99,0
jo	7,83	137,5	63,1	0,84	99,2	**us**	8,10	149,9	68,3	0,68	98,7
jp	7,21	101,0	55,1	0,41	88,1	**uy**	1,74	71,5	52,3	0,67	100,0
ke	3,74	87,6	66,8	0,56	98,2	**uz**	96,1	22,0	0,17	95,6	
kg	8,08	141,4	84,6	0,88	99,0	**vc**	5,18	71,8	72,0	0,35	97,1
kh	4,71	96,3	17,8	0,75	81,5	**ve**	134,1	77,7			
ki	4,31	138,6	38,0	0,25	99,6	**vg**	2,94	147,2	58,1	0,69	95,0
km	4,87	119,5	32,4	0,16	80,5	**vn**	85,9	25,7	0,10	87,5	
kr	50,8	14,6	0,03			**vu**	33,6	0,37	99,1	94,8	0,81
kw	3,41	59,9	8,5	0,02	58,8	**ws**	2,64	53,7	26,7	0,02	54,1
ky	8,11	129,7	95,1	0,87		**ye**	7,79	159,9	56,2	0,65	87,0
kz	3,88	171,6	100,0	0,60	96,1	**za**	5,68	89,2	27,9	0,44	86,7

Table A.8 continued from previous page

TLD	DEM	MOB	NET	SEC	LIT	TLD	DEM	MOB	NET	SEC	LIT
la	81,1	98,9				**zm**	2,64	89,4	27,1	0,19	88,7
lb	3,30	142,3	76,4	0,78	99,8	**zw**		3,32	43,1	14,3	0,10

Table A.9: Top Level Domain and languages used for KenLM

TLD	Languages	TLD	Languages
ad	['catalan']	lr	['english']
ag	['english']	ls	['english']
ao	['portuguese']	lu	['french', 'german']
ar	['spanish']	lv	['latvian']
as	['english']	ly	['english']
at	['german', 'croatian', 'slovenian']	mc	['french']
au	['english', 'french']	me	['english']
ba	['croatian']	mg	['french']
bb	['english', 'french']	mil	['english']
be	['dutch', 'french', 'german']	ml	['english']
bf	['english', 'french']	mm	['english']
bi	['english', 'french']	mr	['english']
bj	['english', 'french']	ms	['english']
bm	['english']	mt	['english']
bn	['english', 'french']	mu	['english']
bo	['spanish']	mv	['english']
br	['portuguese']	mw	['english']
bs	['english', 'french']	mx	['spanish']
bt	['english']	mz	['portuguese']
bw	['english']	na	['english']
by	['english', 'french']	nf	['english']
bz	['english']	ng	['english']
ca	['english']	ni	['spanish']
cd	['french']	nl	['dutch']
cf	['english', 'french']	no	['norwegian']
cg	['english', 'french']	nu	['english']
ci	['french']	nz	['english']
cl	['spanish']	pa	['spanish']
cm	['english', 'french']	pe	['spanish']
co	['spanish']	pf	['english']
cr	['spanish']	pg	['english']
cu	['spanish']	ph	['english']
cv	['portuguese']	pk	['english']
cw	['english', 'french']	pl	['polish']
cz	['czech', 'slovak']	pr	['spanish', 'english']
de	['german']	ps	['spanish', 'english']
dj	['english', 'french']	pt	['portuguese']
dk	['danish']	py	['spanish']
dm	['english', 'french']	ro	['romanian']
do	['spanish']	rs	['spanish', 'english']
ec	['spanish']	rw	['spanish', 'english']
edu	['english']	sb	['spanish', 'english']
ee	['estonian']	sc	['english', 'french']
es	['spanish', 'catalan']	sd	['english']
fi	['finnish', 'swedish']	se	['swedish']

Table A.9 continued from previous page

TLD	Languages
fj	['english']
fm	['english']
fo	['english']
fr	['french']
ga	['english', 'french']
gd	['english', 'french']
gh	['french']
gl	['danish', 'english']
gm	['english', 'french']
gn	['english', 'french']
gov	['english']
gq	['english', 'french']
gt	['spanish']
gw	['english', 'french']
gy	['english', 'french']
hn	['spanish']
hr	['croatian']
ht	['english', 'french']
ch	['french', 'german', 'italian']
ie	['english', 'irish']
it	['italian']
jm	['english']
ky	['english']
lc	['english']
li	['german']

TLD	Languages
sk	['slovak', 'german', 'polish', 'czech']
sl	['spanish', 'english']
sm	['spanish', 'english']
sn	['french']
sr	['spanish', 'english']
ss	['spanish', 'english']
st	['spanish', 'english']
sv	['spanish']
sz	['english']
td	['spanish', 'english']
tg	['spanish', 'english']
tj	['spanish', 'english']
tk	['english']
tl	['spanish', 'english']
to	['english']
tr	['turkish']
tt	['spanish', 'english']
tv	['english']
uk	['english']
us	['english', 'french', 'german', 'spanish', 'italian']
uy	['spanish']
ve	['spanish']
vu	['english', 'french']
ws	['english']
zm	['english']

A.3 Word Break Algorithm

Listing A.1: Word Break Algorithm

```
def parse_pass(password, dictionary, correction=False):
    '''
    Returns all possible splits of a assword given the provided dictionary

    password -  a string
    dictionary - a list of words
    correction - whether to apply filtering correction for too large passwords
    '''

    max_coverage = [0] #how many characters can be covered up to position i
    coverages = [[[]]] #words fromt he dictionary that make the coverage

    plen = len(password)
    mid = math.ceil(plen/2)

    #go iteratively character by character
    for i in range(len(password)):

        #initialize n of covered letters (take previous) and covered words
        # (take previous) and potentionally over write
        max_coverage.append(max_coverage[-1])
        coverages.append([c for c in coverages[-1]])

        #iterate from the beginnign up to i(i is the max of characters
        # included in the dict.) icecream cecream ecream cream ream ream am m
        for zacatek in range(0, i+1):

            #if the moving part form the start to position i is in the dict
            # ('cream' might be, 'ream' might not)
            if password[zacatek:(i+1)] in dictionary:

                #length of the coverage up to the current word
                coverage = i - zacatek + 1 + max_coverage[zacatek]


               #if the new coverage is bigger than previous one
                if coverage > max_coverage[-1]:

                    #just append
                    max_coverage[-1] = coverage
                    coverages[-1] = []# clear space


                #a condition that we have performed an update
                if coverage == max_coverage[-1]:
                        # the condition that we have found something new
                        #  (could be multi in one run)

                    # that is between start and [-2] the coverage did not increase
                    if max_coverage[-2] == max_coverage[zacatek]:
                        #we add a word to previous stuff.
```

```
                    coverages[-1].extend([
                        c + [password[zacatek:(i+1)]] for c in coverages[-2]
                    ])

                #that is between start and [-2] is another word auto mat
                elif max_coverage[-2] > max_coverage[zacatek]:

                    #in this case we are overlapping
                    coverages[-1].extend([
                      c+[password[zacatek:(i+1)]] for c in coverages[zacatek]
                    ])

        #simplify the possibilities for long passwords
        #we expect that an average length of a word is 3 characters while
        # the reality is 4+
        #thus we opt for a safer simplification
        if (correction == True) & (i > 15):

            if i < 20:
                thresh = math.ceil(i/(2.5))
                splitted = [x for x in coverages[-1] if len(x) < thresh]
            elif i < 30:
                thresh = math.ceil(i/(3))
                splitted = [x for x in coverages[-1] if len(x) < thresh]
            elif i < 40:
                thresh = math.ceil(i/(3.5))
                splitted = [x for x in coverages[-1] if len(x) < thresh]
            elif i < 45:
                thresh = math.ceil(i/(4))
                splitted = [x for x in coverages[-1] if len(x) < thresh]
            elif i < 50:
                thresh = math.ceil(i/(4.5))
                splitted = [x for x in coverages[-1] if len(x) < thresh]
            else:
                thresh = math.ceil(i/(5))
                splitted = [x for x in coverages[-1] if len(x) < thresh]


            if len(splitted) == 0:
                thresh = math.ceil(i/2.5)
                splitted = [x for x in coverages[-1] if len(x) < thresh]

            coverages[-1] = splitted
            del splitted


    return coverages[-1]
```

A.4 Model family 1 estimates

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
(Intercept):1	-4.73^{***}	-4.54^{***}	-4.72^{***}	-4.53^{***}	-4.51^{***}	-5.41^{***}
	(0.18)	(0.16)	(0.10)	(0.16)	(0.10)	(0.49)
(Intercept):2	-4.15^{***}	-3.94^{***}	-4.16^{***}	-3.89^{***}	-3.89^{***}	-4.46^{***}
	(0.13)	(0.12)	(0.07)	(0.12)	(0.07)	(0.02)

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
(Intercept):3	-3.57^{***}	-3.42^{***}	-3.66^{***}	-3.30^{***}	-3.37^{***}	-3.98^{***}
	(0.11)	(0.10)	(0.06)	(0.10)	(0.06)	(0.02)
(Intercept):4	-3.00^{***}	-2.92^{***}	-3.17^{***}	-2.83^{***}	-2.87^{***}	-3.49^{***}
	(0.10)	(0.09)	(0.05)	(0.09)	(0.05)	(0.10)
(Intercept):5	-2.29^{***}	-2.29^{***}	-2.51^{***}	-2.19^{***}	-2.24^{***}	-2.86^{***}
	(0.08)	(0.07)	(0.05)	(0.08)	(0.05)	(0.16)
(Intercept):6	-1.50^{***}	-1.57^{***}	-1.77^{***}	-1.42^{***}	-1.51^{***}	-2.20^{***}
	(0.07)	(0.06)	(0.04)	(0.06)	(0.04)	(0.18)
(Intercept):7	-0.74^{***}	-0.89^{***}	-1.04^{***}	-0.72^{***}	-0.80^{***}	-1.64^{***}
	(0.06)	(0.06)	(0.04)	(0.06)	(0.04)	(0.17)
(Intercept):8	-0.11^{*}	-0.28^{***}	-0.40^{***}	-0.18^{***}	-0.17^{***}	-1.13^{***}
	(0.06)	(0.05)	(0.03)	(0.05)	(0.03)	(0.16)
(Intercept):9	0.34^{***}	0.19^{***}	0.08^{**}	0.21^{***}	0.32^{***}	-0.75^{***}
	(0.05)	(0.05)	(0.03)	(0.05)	(0.03)	(0.15)
(Intercept):10	0.66^{***}	0.55^{***}	0.45^{***}	0.47^{***}	0.69^{***}	-0.44^{**}
	(0.05)	(0.05)	(0.03)	(0.04)	(0.03)	(0.15)
(Intercept):11	0.94^{***}	0.83^{***}	0.73^{***}	0.67^{***}	0.98^{***}	-0.26
	(0.05)	(0.05)	(0.03)	(0.04)	(0.03)	(0.15)
(Intercept):12	1.18^{***}	1.07^{***}	0.99^{***}	0.85^{***}	1.26^{***}	-0.10
	(0.05)	(0.05)	(0.03)	(0.04)	(0.03)	(0.15)
(Intercept):13	1.45^{***}	1.32^{***}	1.26^{***}	1.03^{***}	1.53^{***}	0.03
	(0.05)	(0.05)	(0.03)	(0.05)	(0.03)	(0.16)
(Intercept):14	1.77^{***}	1.58^{***}	1.53^{***}	1.23^{***}	1.82^{***}	0.17
	(0.06)	(0.05)	(0.03)	(0.05)	(0.03)	(0.17)
(Intercept):15	2.12^{***}	1.91^{***}	1.88^{***}	1.54^{***}	2.17^{***}	0.36
	(0.07)	(0.06)	(0.04)	(0.05)	(0.04)	(0.19)
(Intercept):16	2.57^{***}	2.28^{***}	2.26^{***}	1.87^{***}	2.57^{***}	0.53^{*}
	(0.08)	(0.07)	(0.04)	(0.06)	(0.04)	(0.22)
(Intercept):17	2.92^{***}	2.67^{***}	2.68^{***}	2.25^{***}	2.98^{***}	0.76^{**}
	(0.09)	(0.08)	(0.05)	(0.07)	(0.05)	(0.26)
(Intercept):18	3.27^{***}	3.04^{***}	3.08^{***}	2.60^{***}	3.37^{***}	1.10^{***}
	(0.10)	(0.09)	(0.05)	(0.08)	(0.06)	(0.32)
(Intercept):19	3.55^{***}	3.34^{***}	3.41^{***}	2.85^{***}	3.69^{***}	1.31^{***}
	(0.12)	(0.10)	(0.06)	(0.09)	(0.07)	(0.37)
(Intercept):20	3.85^{***}	3.63^{***}	3.74^{***}	3.10^{***}	3.99^{***}	1.92^{***}
	(0.13)	(0.12)	(0.07)	(0.11)	(0.08)	(0.51)
(Intercept):21	4.20^{***}	3.93^{***}	4.10^{***}	3.41^{***}	4.28^{***}	2.11^{***}
	(0.16)	(0.14)	(0.08)	(0.12)	(0.09)	(0.57)
(Intercept):22	4.53^{***}	4.31^{***}	4.55^{***}	3.82^{***}	4.64^{***}	3.36^{***}
	(0.19)	(0.16)	(0.10)	(0.15)	(0.10)	(0.88)
(Intercept):23	4.78^{***}	4.55^{***}	4.80^{***}	3.99^{***}	4.91^{***}	4.78^{**}
	(0.22)	(0.18)	(0.12)	(0.17)	(0.12)	(1.53)
(Intercept):24	5.07^{***}	4.83^{***}	5.13^{***}	4.37^{***}	5.21^{***}	5.03^{**}
	(0.25)	(0.21)	(0.14)	(0.21)	(0.14)	(1.90)
(Intercept):25	5.58^{***}	5.27^{***}	5.62^{***}	4.69^{***}	5.65^{***}	
	(0.30)	(0.26)	(0.16)	(0.24)	(0.17)	
(Intercept):26	5.84^{***}	5.59^{***}	6.03^{***}	4.97^{***}	6.04^{***}	
	(0.37)	(0.31)	(0.20)	(0.29)	(0.21)	
(Intercept):27	5.99^{***}	5.92^{***}	6.38^{***}	5.46^{***}	6.32^{***}	
	(0.44)	(0.38)	(0.24)	(0.36)	(0.26)	
(Intercept):28	7.23^{***}	6.74^{***}	7.30^{***}	6.18^{***}	6.99^{***}	
	(0.58)	(0.50)	(0.32)	(0.50)	(0.36)	
(Intercept):29	7.93^{***}	7.75^{***}	8.62^{***}	7.28^{***}	7.49^{***}	
	(0.90)	(0.79)	(0.47)	(0.85)	(0.51)	
PassLen:1	0.12^{***}	0.11^{***}	0.10^{***}	0.13^{***}	0.11^{***}	0.15^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
PassLen:2	0.11^{***}	0.10^{***}	0.10^{***}	0.11^{***}	0.10^{***}	0.15^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
PassLen:3	0.08^{***}	0.07^{***}	0.07^{***}	0.07^{***}	0.07^{***}	0.10^{***}
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
PassLen:4	0.04^{***}	0.04^{***}	0.04^{***}	0.04^{***}	0.04^{***}	0.07^{***}
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
PassLen:5	0.01	0.01	0.01^{*}	0.01^{*}	0.02^{***}	0.03^{***}
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
PassLen:6	-0.03^{***}	-0.02^{***}	-0.03^{***}	-0.03^{***}	-0.02^{***}	-0.02^{***}
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
PassLen:7	-0.06^{***}	-0.05^{***}	-0.05^{***}	-0.06^{***}	-0.05^{***}	-0.05^{***}
	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)
PassLen:8	-0.09^{***}	-0.07^{***}	-0.07^{***}	-0.08^{***}	-0.07^{***}	-0.07^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
PassLen:9	-0.09^{***}	-0.07^{***}	-0.08^{***}	-0.08^{***}	-0.07^{***}	-0.07^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PassLen:10	-0.08^{***}	-0.07^{***}	-0.07^{***}	-0.07^{***}	-0.07^{***}	-0.07^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PassLen:11	-0.07^{***}	-0.06^{***}	-0.06^{***}	-0.05^{***}	-0.05^{***}	-0.05^{***}

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PassLen:12	-0.05^{***}	-0.04^{***}	-0.04^{***}	-0.03^{***}	-0.04^{***}	-0.03^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PassLen:13	-0.03^{***}	-0.02^{***}	-0.02^{***}	-0.01^{*}	-0.02^{***}	-0.01
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
PassLen:14	-0.01^{*}	-0.01	-0.01^{***}	0.01^{*}	-0.00	0.02^{**}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
PassLen:15	0.00	0.00	0.00	0.02^{***}	0.01^{**}	0.03^{***}
	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)
PassLen:16	0.01	0.01	0.01^{*}	0.03^{***}	0.02^{***}	0.04^{***}
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)
PassLen:17	0.01^{*}	0.01	0.01^{**}	0.03^{***}	0.02^{***}	0.05^{***}
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)
PassLen:18	0.02^{*}	0.01	0.01^{***}	0.03^{***}	0.02^{***}	0.05^{***}
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)
PassLen:19	0.03^{***}	0.02^{*}	0.02^{***}	0.04^{***}	0.03^{***}	0.06^{***}
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)
PassLen:20	0.03^{**}	0.02^{*}	0.02^{***}	0.04^{***}	0.03^{***}	0.06^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
PassLen:21	0.03^{**}	0.03^{*}	0.02^{***}	0.04^{***}	0.04^{***}	0.06^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
PassLen:22	0.04^{**}	0.03^{*}	0.03^{***}	0.04^{**}	0.04^{***}	0.06^{**}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
PassLen:23	0.04^{**}	0.04^{*}	0.03^{***}	0.05^{**}	0.04^{***}	0.09^{*}
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.04)
PassLen:24	0.04^{*}	0.04^{*}	0.03^{***}	0.05^{*}	0.04^{***}	0.12^{*}
	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.05)
PassLen:25	0.02	0.03	0.03^{**}	0.05^{*}	0.04^{***}	
	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	
PassLen:26	0.04	0.04	0.04^{**}	0.06^{*}	0.04^{***}	
	(0.02)	(0.02)	(0.01)	(0.03)	(0.01)	
PassLen:27	0.04	0.05	0.05^{**}	0.06	0.05^{***}	
	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	
PassLen:28	0.01	0.03	0.03	0.06	0.08^{***}	
	(0.04)	(0.04)	(0.02)	(0.05)	(0.02)	
PassLen:29	0.00	0.04	0.05	0.06	0.10^{**}	
	(0.06)	(0.06)	(0.03)	(0.08)	(0.03)	
Cyber:1	0.38^{**}	0.33^{*}	0.55^{***}	-0.15	0.17	0.09
	(0.15)	(0.15)	(0.11)	(0.15)	(0.10)	(0.53)
Cyber:2	0.43^{***}	0.33^{**}	0.63^{***}	-0.16	0.23^{**}	0.09^{***}
	(0.11)	(0.12)	(0.08)	(0.11)	(0.07)	(0.00)
Cyber:3	0.52^{***}	0.42^{***}	0.75^{***}	-0.15	0.30^{***}	0.24^{***}
	(0.09)	(0.10)	(0.07)	(0.10)	(0.06)	(0.00)
Cyber:4	0.56^{***}	0.47^{***}	0.84^{***}	-0.14	0.34^{***}	0.31^{**}
	(0.08)	(0.09)	(0.06)	(0.09)	(0.05)	(0.10)
Cyber:5	0.47^{***}	0.36^{***}	0.69^{***}	-0.15^{*}	0.27^{***}	0.35^{*}
	(0.07)	(0.07)	(0.05)	(0.07)	(0.04)	(0.17)
Cyber:6	0.33^{***}	0.25^{***}	0.55^{***}	-0.18^{**}	0.15^{***}	0.48^{*}
	(0.06)	(0.06)	(0.04)	(0.06)	(0.04)	(0.19)
Cyber:7	0.13^{*}	0.09	0.33^{***}	-0.27^{***}	-0.01	0.50^{**}
	(0.05)	(0.06)	(0.04)	(0.05)	(0.03)	(0.18)
Cyber:8	-0.07	-0.08	0.12^{***}	-0.35^{***}	-0.20^{***}	0.50^{**}
	(0.05)	(0.05)	(0.03)	(0.05)	(0.03)	(0.17)
Cyber:9	-0.24^{***}	-0.23^{***}	-0.06	-0.39^{***}	-0.37^{***}	0.49^{**}
	(0.04)	(0.05)	(0.03)	(0.04)	(0.03)	(0.16)
Cyber:10	-0.39^{***}	-0.36^{***}	-0.21^{***}	-0.43^{***}	-0.50^{***}	0.45^{**}
	(0.04)	(0.05)	(0.03)	(0.04)	(0.03)	(0.16)
Cyber:11	-0.51^{***}	-0.47^{***}	-0.34^{***}	-0.46^{***}	-0.64^{***}	0.45^{**}
	(0.04)	(0.05)	(0.03)	(0.04)	(0.03)	(0.16)
Cyber:12	-0.60^{***}	-0.56^{***}	-0.46^{***}	-0.51^{***}	-0.76^{***}	0.45^{**}
	(0.04)	(0.05)	(0.03)	(0.04)	(0.03)	(0.16)
Cyber:13	-0.70^{***}	-0.68^{***}	-0.60^{***}	-0.59^{***}	-0.90^{***}	0.41^{*}
	(0.04)	(0.05)	(0.03)	(0.04)	(0.03)	(0.17)
Cyber:14	-0.81^{***}	-0.78^{***}	-0.72^{***}	-0.67^{***}	-1.06^{***}	0.39^{*}
	(0.05)	(0.05)	(0.04)	(0.05)	(0.03)	(0.19)
Cyber:15	-0.96^{***}	-0.94^{***}	-0.91^{***}	-0.80^{***}	-1.23^{***}	0.39
	(0.05)	(0.06)	(0.04)	(0.05)	(0.04)	(0.21)
Cyber:16	-1.11^{***}	-1.09^{***}	-1.07^{***}	-0.88^{***}	-1.42^{***}	0.48^{*}
	(0.06)	(0.07)	(0.05)	(0.06)	(0.04)	(0.24)
Cyber:17	-1.26^{***}	-1.24^{***}	-1.28^{***}	-0.99^{***}	-1.58^{***}	0.52
	(0.07)	(0.08)	(0.05)	(0.07)	(0.05)	(0.29)
Cyber:18	-1.39^{***}	-1.37^{***}	-1.45^{***}	-1.02^{***}	-1.71^{***}	0.52
	(0.08)	(0.09)	(0.06)	(0.08)	(0.06)	(0.35)
Cyber:19	-1.40^{***}	-1.41^{***}	-1.55^{***}	-1.02^{***}	-1.77^{***}	0.52
	(0.10)	(0.10)	(0.07)	(0.09)	(0.06)	(0.41)

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
Cyber:20	-1.47^{***}	-1.44^{***}	-1.64^{***}	-1.01^{***}	-1.81^{***}	0.52
	(0.11)	(0.12)	(0.08)	(0.10)	(0.08)	(0.57)
Cyber:21	-1.53^{***}	-1.47^{***}	-1.74^{***}	-0.98^{***}	-1.84^{***}	0.42
	(0.14)	(0.14)	(0.10)	(0.12)	(0.09)	(0.64)
Cyber:22	-1.72^{***}	-1.58^{***}	-1.97^{***}	-0.98^{***}	-1.91^{***}	-0.09
	(0.16)	(0.16)	(0.11)	(0.15)	(0.11)	(0.99)
Cyber:23	-1.71^{***}	-1.60^{***}	-1.99^{***}	-0.97^{***}	-1.91^{***}	-0.88
	(0.19)	(0.19)	(0.13)	(0.17)	(0.12)	(1.75)
Cyber:24	-1.76^{***}	-1.59^{***}	-2.04^{***}	-0.99^{***}	-1.91^{***}	-0.88
	(0.22)	(0.22)	(0.16)	(0.21)	(0.15)	(2.16)
Cyber:25	-1.81^{***}	-1.74^{***}	-2.26^{***}	-1.03^{***}	-1.99^{***}	
	(0.27)	(0.26)	(0.19)	(0.24)	(0.18)	
Cyber:26	-1.82^{***}	-1.80^{***}	-2.45^{***}	-1.03^{***}	-2.15^{***}	
	(0.33)	(0.32)	(0.23)	(0.29)	(0.22)	
Cyber:27	-1.71^{***}	-1.81^{***}	-2.47^{***}	-1.04^{**}	-2.13^{***}	
	(0.39)	(0.39)	(0.28)	(0.36)	(0.27)	
Cyber:28	-1.84^{***}	-2.04^{***}	-2.84^{***}	-1.21^{*}	-2.42^{***}	
	(0.51)	(0.52)	(0.37)	(0.50)	(0.38)	
Cyber:29	-2.39^{**}	-2.45^{**}	-3.75^{***}	-1.30	-2.39^{***}	
	(0.83)	(0.84)	(0.56)	(0.86)	(0.53)	
Mobile:1	-0.01					
	(0.12)					
Mobile:2	-0.06					
	(0.09)					
Mobile:3	-0.08					
	(0.08)					
Mobile:4	-0.10					
	(0.07)					
Mobile:5	-0.14^{*}					
	(0.06)					
Mobile:6	-0.13^{**}					
	(0.05)					
Mobile:7	-0.11^{*}					
	(0.04)					
Mobile:8	-0.06					
	(0.04)					
Mobile:9	-0.02					
	(0.04)					
Mobile:10	0.00					
	(0.04)					
Mobile:11	0.00					
	(0.03)					
Mobile:12	0.00					
	(0.04)					
Mobile:13	-0.02					
	(0.04)					
Mobile:14	-0.07					
	(0.04)					
Mobile:15	-0.08					
	(0.05)					
Mobile:16	-0.12^{*}					
	(0.05)					
Mobile:17	-0.09					
	(0.06)					
Mobile:18	-0.06					
	(0.07)					
Mobile:19	-0.08					
	(0.08)					
Mobile:20	-0.02					
	(0.09)					
Mobile:21	-0.01					
	(0.11)					
Mobile:22	0.08					
	(0.13)					
Mobile:23	0.07					
	(0.15)					
Mobile:24	0.10					
	(0.17)					
Mobile:25	0.06					
	(0.20)					
Mobile:26	0.09					
	(0.25)					
Mobile:27	0.22					
	(0.31)					
Mobile:28	-0.08					

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
	(0.38)					
Mobile:29	0.25					
	(0.61)					
Effort2:1	-0.68^{***}	-0.58^{***}	-0.61^{***}	-0.79^{***}	-0.56^{***}	-0.38^{***}
	(0.07)	(0.08)	(0.04)	(0.09)	(0.04)	(0.08)
Effort2:2	-0.55^{***}	-0.50^{***}	-0.51^{***}	-0.65^{***}	-0.47^{***}	-0.38^{***}
	(0.05)	(0.06)	(0.03)	(0.06)	(0.03)	(0.03)
Effort2:3	-0.30^{***}	-0.26^{***}	-0.27^{***}	-0.32^{***}	-0.26^{***}	-0.19^{***}
	(0.04)	(0.05)	(0.03)	(0.05)	(0.03)	(0.00)
Effort2:4	-0.14^{***}	-0.13^{**}	-0.13^{***}	-0.17^{***}	-0.13^{***}	-0.06^{***}
	(0.04)	(0.04)	(0.02)	(0.05)	(0.02)	(0.00)
Effort2:5	-0.04	-0.04	-0.04	-0.05	-0.04^{*}	-0.01
	(0.03)	(0.04)	(0.02)	(0.04)	(0.02)	(0.02)
Effort2:6	-0.03	-0.02	-0.02	-0.03	-0.02	0.01
	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)
Effort2:7	0.00	-0.00	-0.00	-0.02	0.00	0.01
	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)
Effort2:8	0.03	0.02	0.02	-0.01	0.02	0.02
	(0.02)	(0.03)	(0.01)	(0.03)	(0.01)	(0.03)
Effort2:9	0.04	0.03	0.03^{*}	0.00	0.03^{*}	0.02
	(0.02)	(0.03)	(0.01)	(0.02)	(0.01)	(0.03)
Effort2:10	0.04^{*}	0.03	0.03^{*}	0.00	0.03^{*}	0.02
	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.03)
Effort2:11	0.03	0.03	0.03^{*}	0.00	0.03^{*}	-0.00
	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.03)
Effort2:12	0.05^{*}	0.03	0.03^{*}	0.00	0.03^{*}	-0.01
	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.03)
Effort2:13	0.04	0.02	0.02	0.00	0.02	-0.02
	(0.02)	(0.03)	(0.01)	(0.02)	(0.01)	(0.03)
Effort2:14	0.04	0.02	0.02	0.00	0.02	-0.03
	(0.02)	(0.03)	(0.01)	(0.03)	(0.02)	(0.03)
Effort2:15	0.03	0.02	0.02	0.01	0.02	-0.04
	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)
Effort2:16	0.01	0.02	0.03	0.03	0.02	-0.05
	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.04)
Effort2:17	0.03	0.03	0.03	0.04	0.03	-0.05
	(0.03)	(0.04)	(0.02)	(0.04)	(0.02)	(0.04)
Effort2:18	0.05	0.03	0.04	0.07	0.03	-0.03
	(0.04)	(0.04)	(0.02)	(0.04)	(0.02)	(0.05)
Effort2:19	0.05	0.03	0.04	0.09^{*}	0.03	-0.02
	(0.04)	(0.05)	(0.03)	(0.05)	(0.03)	(0.06)
Effort2:20	0.04	0.03	0.03	0.13^{*}	0.03	-0.02
	(0.05)	(0.05)	(0.03)	(0.05)	(0.03)	(0.08)
Effort2:21	0.03	0.03	0.03	0.13^{*}	0.03	-0.05
	(0.06)	(0.06)	(0.03)	(0.06)	(0.03)	(0.08)
Effort2:22	0.01	0.03	0.04	0.13	0.03	-0.07
	(0.07)	(0.07)	(0.04)	(0.08)	(0.04)	(0.12)
Effort2:23	0.06	0.07	0.07	0.17^{*}	0.03	-0.07
	(0.08)	(0.08)	(0.05)	(0.08)	(0.05)	(0.20)
Effort2:24	0.12	0.10	0.10	0.22^{*}	0.03	-0.07
	(0.09)	(0.10)	(0.05)	(0.10)	(0.05)	(0.25)
Effort2:25	0.16	0.14	0.12^{*}	0.23	0.06	
	(0.10)	(0.11)	(0.06)	(0.12)	(0.06)	
Effort2:26	0.20	0.16	0.14	0.23	0.11	
	(0.13)	(0.14)	(0.07)	(0.14)	(0.08)	
Effort2:27	0.17	0.19	0.18	0.22	0.12	
	(0.16)	(0.17)	(0.09)	(0.18)	(0.09)	
Effort2:28	0.12	0.17	0.18	0.20	0.10	
	(0.20)	(0.21)	(0.12)	(0.25)	(0.13)	
Effort2:29	0.37	0.33	0.35^{*}	0.23	0.05	
	(0.31)	(0.33)	(0.17)	(0.42)	(0.18)	
Effort3:1	-1.48^{***}	-1.51^{***}	-1.54^{***}	-2.50^{***}	-1.60^{***}	-1.98^{***}
	(0.14)	(0.16)	(0.09)	(0.26)	(0.09)	(0.26)
Effort3:2	-1.40^{***}	-1.47^{***}	-1.49^{***}	-2.10^{***}	-1.57^{***}	-1.74^{***}
	(0.10)	(0.12)	(0.06)	(0.16)	(0.07)	(0.14)
Effort3:3	-1.24^{***}	-1.29^{***}	-1.30^{***}	-1.83^{***}	-1.34^{***}	-1.52^{***}
	(0.08)	(0.10)	(0.05)	(0.13)	(0.05)	(0.12)
Effort3:4	-1.03^{***}	-1.05^{***}	-1.06^{***}	-1.51^{***}	-1.12^{***}	-1.27^{***}
	(0.07)	(0.08)	(0.04)	(0.10)	(0.04)	(0.10)
Effort3:5	-0.95^{***}	-0.94^{***}	-0.95^{***}	-1.37^{***}	-1.00^{***}	-1.19^{***}
	(0.06)	(0.07)	(0.04)	(0.08)	(0.04)	(0.08)
Effort3:6	-0.82^{***}	-0.81^{***}	-0.82^{***}	-1.10^{***}	-0.84^{***}	-1.07^{***}
	(0.05)	(0.05)	(0.03)	(0.06)	(0.03)	(0.07)
Effort3:7	-0.72^{***}	-0.70^{***}	-0.70^{***}	-0.90^{***}	-0.72^{***}	-0.86^{***}
	(0.04)	(0.05)	(0.03)	(0.05)	(0.03)	(0.06)

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
Effort3:8	-0.57^{***}	-0.55^{***}	-0.55^{***}	-0.67^{***}	-0.57^{***}	-0.67^{***}
	(0.04)	(0.04)	(0.02)	(0.04)	(0.02)	(0.05)
Effort3:9	-0.42^{***}	-0.42^{***}	-0.41^{***}	-0.50^{***}	-0.42^{***}	-0.49^{***}
	(0.03)	(0.04)	(0.02)	(0.04)	(0.02)	(0.04)
Effort3:10	-0.31^{***}	-0.32^{***}	-0.32^{***}	-0.38^{***}	-0.32^{***}	-0.40^{***}
	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.04)
Effort3:11	-0.23^{***}	-0.24^{***}	-0.23^{***}	-0.29^{***}	-0.24^{***}	-0.31^{***}
	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.04)
Effort3:12	-0.17^{***}	-0.17^{***}	-0.17^{***}	-0.21^{***}	-0.18^{***}	-0.24^{***}
	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.04)
Effort3:13	-0.12^{***}	-0.11^{**}	-0.11^{***}	-0.13^{***}	-0.12^{***}	-0.16^{***}
	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.04)
Effort3:14	-0.08^{**}	-0.05	-0.05^{*}	-0.05	-0.06^{**}	-0.10^{*}
	(0.03)	(0.04)	(0.02)	(0.03)	(0.02)	(0.04)
Effort3:15	-0.02	0.02	0.02	0.04	0.01	-0.01
	(0.03)	(0.04)	(0.02)	(0.04)	(0.02)	(0.05)
Effort3:16	0.02	0.09	0.09^{***}	0.11^{*}	0.08^{***}	0.05
	(0.04)	(0.04)	(0.02)	(0.04)	(0.02)	(0.05)
Effort3:17	0.09^{*}	0.16^{**}	0.15^{***}	0.20^{***}	0.14^{***}	0.13^{*}
	(0.04)	(0.05)	(0.03)	(0.05)	(0.03)	(0.06)
Effort3:18	0.15^{**}	0.23^{***}	0.22^{***}	0.26^{***}	0.21^{***}	0.18^{*}
	(0.05)	(0.06)	(0.03)	(0.06)	(0.03)	(0.07)
Effort3:19	0.22^{***}	0.27^{***}	0.27^{***}	0.31^{***}	0.27^{***}	0.22^{**}
	(0.06)	(0.07)	(0.04)	(0.07)	(0.04)	(0.08)
Effort3:20	0.27^{***}	0.35^{***}	0.35^{***}	0.37^{***}	0.30^{***}	0.32^{**}
	(0.07)	(0.08)	(0.04)	(0.08)	(0.04)	(0.12)
Effort3:21	0.35^{***}	0.41^{***}	0.41^{***}	0.42^{***}	0.37^{***}	0.33^{**}
	(0.09)	(0.10)	(0.05)	(0.09)	(0.05)	(0.13)
Effort3:22	0.43^{***}	0.46^{***}	0.48^{***}	0.49^{***}	0.47^{***}	0.33
	(0.10)	(0.11)	(0.06)	(0.12)	(0.06)	(0.19)
Effort3:23	0.57^{***}	0.55^{***}	0.57^{***}	0.61^{***}	0.54^{***}	0.33
	(0.13)	(0.13)	(0.07)	(0.14)	(0.07)	(0.31)
Effort3:24	0.72^{***}	0.64^{***}	0.67^{***}	0.67^{***}	0.61^{***}	0.33
	(0.16)	(0.16)	(0.09)	(0.17)	(0.09)	(0.39)
Effort3:25	0.92^{***}	0.82^{***}	0.86^{***}	0.92^{***}	0.76^{***}	
	(0.20)	(0.20)	(0.11)	(0.22)	(0.11)	
Effort3:26	0.89^{***}	0.93^{***}	0.93^{***}	1.32^{***}	0.89^{***}	
	(0.24)	(0.25)	(0.14)	(0.31)	(0.14)	
Effort3:27	0.94^{**}	1.07^{***}	1.05^{***}	1.34^{***}	1.06^{***}	
	(0.30)	(0.32)	(0.17)	(0.40)	(0.18)	
Effort3:28	1.06^{**}	1.28^{**}	1.27^{***}	1.41^{*}	1.04^{***}	
	(0.40)	(0.45)	(0.24)	(0.57)	(0.25)	
Effort3:29	1.19^{*}	1.33^{*}	1.37^{***}	1.65	1.28^{**}	
	(0.61)	(0.67)	(0.37)	(1.05)	(0.40)	
Effort4:1	-5.26^{**}	-4.83^{**}	-4.90^{***}	-14.87	-4.95^{***}	-3.58^{**}
	(1.95)	(1.76)	(0.98)	(310.27)	(1.01)	(1.20)
Effort4:2	-4.67^{***}	-4.18^{***}	-4.10^{***}	-14.87	-3.91^{***}	-3.58^{***}
	(1.13)	(0.96)	(0.50)	(239.79)	(0.45)	(0.00)
Effort4:3	-3.44^{***}	-3.14^{***}	-3.09^{***}	-7.64	-2.80^{***}	-3.58^{***}
	(0.54)	(0.52)	(0.27)	(5.84)	(0.23)	(0.00)
Effort4:4	-2.56^{***}	-2.41^{***}	-2.39^{***}	-4.23^{***}	-2.23^{***}	-3.58^{***}
	(0.31)	(0.31)	(0.17)	(1.04)	(0.15)	(0.00)
Effort4:5	-1.95^{***}	-2.00^{***}	-1.94^{***}	-2.59^{***}	-1.89^{***}	-3.04^{***}
	(0.20)	(0.22)	(0.12)	(0.38)	(0.11)	(0.43)
Effort4:6	-1.50^{***}	-1.54^{***}	-1.49^{***}	-2.08^{***}	-1.56^{***}	-2.66^{***}
	(0.14)	(0.15)	(0.08)	(0.25)	(0.08)	(0.36)
Effort4:7	-1.19^{***}	-1.22^{***}	-1.21^{***}	-1.72^{***}	-1.29^{***}	-1.93^{***}
	(0.10)	(0.12)	(0.06)	(0.18)	(0.07)	(0.24)
Effort4:8	-0.89^{***}	-0.94^{***}	-0.93^{***}	-1.43^{***}	-0.94^{***}	-1.31^{***}
	(0.08)	(0.09)	(0.05)	(0.14)	(0.05)	(0.16)
Effort4:9	-0.63^{***}	-0.66^{***}	-0.65^{***}	-0.89^{***}	-0.66^{***}	-0.83^{***}
	(0.07)	(0.08)	(0.04)	(0.10)	(0.04)	(0.12)
Effort4:10	-0.44^{***}	-0.45^{***}	-0.43^{***}	-0.55^{***}	-0.48^{***}	-0.56^{***}
	(0.06)	(0.07)	(0.04)	(0.08)	(0.04)	(0.10)
Effort4:11	-0.30^{***}	-0.30^{***}	-0.29^{***}	-0.32^{***}	-0.36^{***}	-0.40^{***}
	(0.06)	(0.06)	(0.04)	(0.07)	(0.04)	(0.09)
Effort4:12	-0.21^{***}	-0.17^{**}	-0.17^{***}	-0.16^{*}	-0.23^{***}	-0.22^{*}
	(0.06)	(0.06)	(0.03)	(0.07)	(0.04)	(0.09)
Effort4:13	-0.11	-0.06	-0.05	-0.09	-0.09^{*}	-0.07
	(0.06)	(0.07)	(0.04)	(0.08)	(0.04)	(0.09)
Effort4:14	-0.03	0.06	0.06	0.18^{*}	0.03	0.14
	(0.06)	(0.07)	(0.04)	(0.08)	(0.04)	(0.11)
Effort4:15	0.03	0.15	0.16^{***}	0.34^{***}	0.13^{**}	0.38^{**}
	(0.07)	(0.08)	(0.04)	(0.10)	(0.04)	(0.13)
Effort4:16	0.10	0.22^{*}	0.24^{***}	0.45^{***}	0.22^{***}	0.43^{**}

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
	(0.08)	(0.09)	(0.05)	(0.12)	(0.05)	(0.15)
Effort4:17	0.19*	0.30**	0.33***	0.65***	0.27***	0.54**
	(0.09)	(0.10)	(0.06)	(0.14)	(0.06)	(0.18)
Effort4:18	0.24*	0.40**	0.41***	0.83***	0.35***	0.65**
	(0.11)	(0.12)	(0.07)	(0.18)	(0.07)	(0.23)
Effort4:19	0.39**	0.54***	0.53***	1.01***	0.44***	0.89**
	(0.13)	(0.15)	(0.08)	(0.22)	(0.08)	(0.28)
Effort4:20	0.61***	0.68***	0.70***	1.39***	0.59***	1.13**
	(0.17)	(0.18)	(0.10)	(0.31)	(0.10)	(0.43)
Effort4:21	0.74***	0.77***	0.79***	1.43***	0.75***	1.13*
	(0.21)	(0.22)	(0.12)	(0.37)	(0.12)	(0.47)
Effort4:22	0.90***	0.86**	0.89***	1.51**	0.88***	1.13
	(0.27)	(0.27)	(0.15)	(0.48)	(0.15)	(0.70)
Effort4:23	0.99**	1.02**	1.07***	1.50**	1.00***	1.13
	(0.33)	(0.34)	(0.18)	(0.53)	(0.18)	(1.16)
Effort4:24	1.57**	1.59**	1.66***	1.76*	1.49***	1.13
	(0.50)	(0.52)	(0.28)	(0.70)	(0.26)	(1.47)
Effort4:25	2.07**	2.03**	2.19***	1.73*	1.77***	
	(0.72)	(0.74)	(0.42)	(0.83)	(0.36)	
Effort4:26	2.19*	2.13*	2.24***	1.69	1.94***	
	(0.94)	(0.94)	(0.52)	(0.98)	(0.46)	
Effort4:27	2.24	2.39	2.55***	1.22	1.88***	
	(1.20)	(1.27)	(0.73)	(0.98)	(0.54)	
Effort4:28	2.47	2.68	2.77**	1.13	1.54*	
	(1.69)	(1.86)	(1.03)	(1.34)	(0.63)	
Effort4:29	5.23	9.08	16.78	1.36	1.75	
	(8.95)	(64.06)	(1579.37)	(2.25)	(1.01)	
SexF:1	−0.01					
	(0.06)					
SexF:2	0.01					
	(0.05)					
SexF:3	−0.02					
	(0.04)					
SexF:4	−0.03					
	(0.03)					
SexF:5	−0.03					
	(0.03)					
SexF:6	−0.03					
	(0.03)					
SexF:7	−0.01					
	(0.02)					
SexF:8	−0.01					
	(0.02)					
SexF:9	0.00					
	(0.02)					
SexF:10	0.01					
	(0.02)					
SexF:11	0.02					
	(0.02)					
SexF:12	0.01					
	(0.02)					
SexF:13	0.01					
	(0.02)					
SexF:14	−0.01					
	(0.02)					
SexF:15	−0.02					
	(0.02)					
SexF:16	−0.02					
	(0.03)					
SexF:17	−0.02					
	(0.03)					
SexF:18	−0.02					
	(0.03)					
SexF:19	−0.02					
	(0.04)					
SexF:20	−0.02					
	(0.04)					
SexF:21	−0.05					
	(0.05)					
SexF:22	−0.04					
	(0.06)					
SexF:23	−0.05					
	(0.07)					
SexF:24	−0.00					
	(0.08)					

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
SexF:25	−0.02					
	(0.10)					
SexF:26	−0.03					
	(0.12)					
SexF:27	−0.04					
	(0.15)					
SexF:28	−0.04					
	(0.19)					
SexF:29	0.27					
	(0.29)					
pca1:1			0.06^{***}			
			(0.01)			
pca1:2			0.06^{***}			
			(0.01)			
pca1:3			0.07^{***}			
			(0.01)			
pca1:4			0.07^{***}			
			(0.01)			
pca1:5			0.06^{***}			
			(0.01)			
pca1:6			0.06^{***}			
			(0.00)			
pca1:7			0.05^{***}			
			(0.00)			
pca1:8			0.04^{***}			
			(0.00)			
pca1:9			0.04^{***}			
			(0.00)			
pca1:10			0.04^{***}			
			(0.00)			
pca1:11			0.04^{***}			
			(0.00)			
pca1:12			0.03^{***}			
			(0.00)			
pca1:13			0.03^{***}			
			(0.00)			
pca1:14			0.03^{***}			
			(0.00)			
pca1:15			0.03^{***}			
			(0.00)			
pca1:16			0.03^{***}			
			(0.00)			
pca1:17			0.03^{***}			
			(0.01)			
pca1:18			0.02^{***}			
			(0.01)			
pca1:19			0.02^{**}			
			(0.01)			
pca1:20			0.02^{**}			
			(0.01)			
pca1:21			0.02^{*}			
			(0.01)			
pca1:22			0.01			
			(0.01)			
pca1:23			0.01			
			(0.01)			
pca1:24			0.01			
			(0.01)			
pca1:25			0.01			
			(0.02)			
pca1:26			−0.01			
			(0.02)			
pca1:27			−0.01			
			(0.02)			
pca1:28			−0.04			
			(0.03)			
pca1:29			-0.10^{*}			
			(0.04)			
pca2:1			−0.01			
			(0.10)			
pca2:2			0.18^{*}			
			(0.08)			
pca2:3			0.23^{***}			
			(0.07)			
pca2:4			0.29^{***}			

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
			(0.06)			
pca2:5			0.25^{***}			
			(0.05)			
pca2:6			0.23^{***}			
			(0.04)			
pca2:7			0.17^{***}			
			(0.04)			
pca2:8			0.11^{**}			
			(0.03)			
pca2:9			0.05			
			(0.03)			
pca2:10			−0.01			
			(0.03)			
pca2:11			−0.04			
			(0.03)			
pca2:12			-0.08^{*}			
			(0.03)			
pca2:13			-0.12^{***}			
			(0.03)			
pca2:14			-0.13^{***}			
			(0.04)			
pca2:15			-0.18^{***}			
			(0.04)			
pca2:16			-0.21^{***}			
			(0.04)			
pca2:17			-0.29^{***}			
			(0.05)			
pca2:18			-0.37^{***}			
			(0.06)			
pca2:19			-0.50^{***}			
			(0.06)			
pca2:20			-0.61^{***}			
			(0.07)			
pca2:21			-0.73^{***}			
			(0.09)			
pca2:22			-0.86^{***}			
			(0.10)			
pca2:23			-0.87^{***}			
			(0.11)			
pca2:24			-0.91^{***}			
			(0.13)			
pca2:25			-0.93^{***}			
			(0.16)			
pca2:26			-1.06^{***}			
			(0.19)			
pca2:27			-1.10^{***}			
			(0.23)			
pca2:28			-1.12^{***}			
			(0.29)			
pca2:29			-1.31^{**}			
			(0.41)			
pca3:1			0.74^{**}			
			(0.26)			
pca3:2			0.50^{**}			
			(0.19)			
pca3:3			0.52^{**}			
			(0.16)			
pca3:4			0.51^{***}			
			(0.14)			
pca3:5			0.47^{***}			
			(0.12)			
pca3:6			0.33^{**}			
			(0.11)			
pca3:7			0.16			
			(0.09)			
pca3:8			0.00			
			(0.09)			
pca3:9			−0.10			
			(0.08)			
pca3:10			-0.22^{**}			
			(0.08)			
pca3:11			-0.34^{***}			
			(0.08)			
pca3:12			-0.49^{***}			
			(0.08)			

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
pca3:13			-0.63^{***}			
			(0.08)			
pca3:14			-0.83^{***}			
			(0.09)			
pca3:15			-0.99^{***}			
			(0.10)			
pca3:16			-1.12^{***}			
			(0.12)			
pca3:17			-1.25^{***}			
			(0.13)			
pca3:18			-1.36^{***}			
			(0.15)			
pca3:19			-1.42^{***}			
			(0.17)			
pca3:20			-1.30^{***}			
			(0.20)			
pca3:21			-1.10^{***}			
			(0.23)			
pca3:22			-0.91^{***}			
			(0.27)			
pca3:23			-0.63^{*}			
			(0.30)			
pca3:24			-0.45			
			(0.35)			
pca3:25			-0.34			
			(0.42)			
pca3:26			-0.42			
			(0.50)			
pca3:27			-0.20			
			(0.60)			
pca3:28			0.07			
			(0.78)			
pca3:29			1.94^{*}			
			(0.94)			
SentPos:1				-0.85^{***}		
				(0.21)		
SentPos:2				-0.87^{***}		
				(0.16)		
SentPos:3				-0.88^{***}		
				(0.14)		
SentPos:4				-0.83^{***}		
				(0.12)		
SentPos:5				-0.83^{***}		
				(0.10)		
SentPos:6				-0.62^{***}		
				(0.07)		
SentPos:7				-0.39^{***}		
				(0.06)		
SentPos:8				-0.12^{*}		
				(0.05)		
SentPos:9				0.03		
				(0.04)		
SentPos:10				0.15^{***}		
				(0.04)		
SentPos:11				0.29^{***}		
				(0.04)		
SentPos:12				0.45^{***}		
				(0.04)		
SentPos:13				0.55^{***}		
				(0.05)		
SentPos:14				0.67^{***}		
				(0.05)		
SentPos:15				0.81^{***}		
				(0.06)		
SentPos:16				0.89^{***}		
				(0.08)		
SentPos:17				1.00^{***}		
				(0.09)		
SentPos:18				1.03^{***}		
				(0.11)		
SentPos:19				1.18^{***}		
				(0.13)		
SentPos:20				1.29^{***}		
				(0.16)		
SentPos:21				1.42^{***}		

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
				(0.20)		
SentPos:22				1.59***		
				(0.27)		
SentPos:23				1.67***		
				(0.31)		
SentPos:24				1.68***		
				(0.39)		
SentPos:25				2.18***		
				(0.58)		
SentPos:26				14.89		
				(368.76)		
SentPos:27				15.43		
				(612.74)		
SentPos:28				15.85		
				(1032.92)		
SentPos:29				15.86		
				(1742.05)		
SentNeg:1				−0.62		
				(0.36)		
SentNeg:2				−0.84**		
				(0.28)		
SentNeg:3				−0.93***		
				(0.28)		
SentNeg:4				−0.99***		
				(0.25)		
SentNeg:5				−0.99***		
				(0.21)		
SentNeg:6				−0.72***		
				(0.15)		
SentNeg:7				−0.48***		
				(0.12)		
SentNeg:8				−0.30**		
				(0.10)		
SentNeg:9				−0.17*		
				(0.09)		
SentNeg:10				−0.06		
				(0.08)		
SentNeg:11				0.13		
				(0.08)		
SentNeg:12				0.27***		
				(0.08)		
SentNeg:13				0.33***		
				(0.09)		
SentNeg:14				0.53***		
				(0.10)		
SentNeg:15				0.69***		
				(0.12)		
SentNeg:16				0.76***		
				(0.14)		
SentNeg:17				0.80***		
				(0.16)		
SentNeg:18				0.96***		
				(0.20)		
SentNeg:19				0.99***		
				(0.24)		
SentNeg:20				1.07***		
				(0.28)		
SentNeg:21				1.26***		
				(0.36)		
SentNeg:22				1.53**		
				(0.50)		
SentNeg:23				1.75**		
				(0.65)		
SentNeg:24				1.97*		
				(0.85)		
SentNeg:25				1.93		
				(1.00)		
SentNeg:26				1.88		
				(1.17)		
SentNeg:27				15.44		
				(1199.65)		
SentNeg:28				15.87		
				(2022.76)		
SentNeg:29				15.87		
				(3409.38)		

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
austro_asiatic:1					0.19	
					(0.13)	
austro_asiatic:2					0.08	
					(0.10)	
austro_asiatic:3					0.10	
					(0.08)	
austro_asiatic:4					0.10	
					(0.07)	
austro_asiatic:5					0.07	
					(0.06)	
austro_asiatic:6					−0.01	
					(0.06)	
austro_asiatic:7					−0.08	
					(0.05)	
austro_asiatic:8					−0.14**	
					(0.05)	
austro_asiatic:9					−0.22***	
					(0.05)	
austro_asiatic:10					−0.27***	
					(0.04)	
austro_asiatic:11					−0.35***	
					(0.04)	
austro_asiatic:12					−0.42***	
					(0.04)	
austro_asiatic:13					−0.51***	
					(0.04)	
austro_asiatic:14					−0.63***	
					(0.04)	
austro_asiatic:15					−0.74***	
					(0.04)	
austro_asiatic:16					−0.87***	
					(0.05)	
austro_asiatic:17					−1.00***	
					(0.05)	
austro_asiatic:18					−1.13***	
					(0.05)	
austro_asiatic:19					−1.22***	
					(0.06)	
austro_asiatic:20					−1.32***	
					(0.06)	
austro_asiatic:21					−1.43***	
					(0.07)	
austro_asiatic:22					−1.52***	
					(0.08)	
austro_asiatic:23					−1.61***	
					(0.08)	
austro_asiatic:24					−1.67***	
					(0.09)	
austro_asiatic:25					−1.78***	
					(0.11)	
austro_asiatic:26					−1.87***	
					(0.13)	
austro_asiatic:27					−1.92***	
					(0.15)	
austro_asiatic:28					−1.97***	
					(0.20)	
austro_asiatic:29					−1.94***	
					(0.28)	
chinese:1					0.60***	
					(0.11)	
chinese:2					0.49***	
					(0.09)	
chinese:3					0.48***	
					(0.08)	
chinese:4					0.48***	
					(0.07)	
chinese:5					0.47***	
					(0.06)	
chinese:6					0.45***	
					(0.05)	
chinese:7					0.43***	
					(0.05)	
chinese:8					0.41***	
					(0.05)	
chinese:9					0.40***	

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
					(0.05)	
chinese:10					0.39***	
					(0.04)	
chinese:11					0.39***	
					(0.05)	
chinese:12					0.38***	
					(0.05)	
chinese:13					0.39***	
					(0.05)	
chinese:14					0.41***	
					(0.05)	
chinese:15					0.44***	
					(0.06)	
chinese:16					0.49***	
					(0.07)	
chinese:17					0.57***	
					(0.08)	
chinese:18					0.61***	
					(0.09)	
chinese:19					0.71***	
					(0.11)	
chinese:20					0.86***	
					(0.14)	
chinese:21					0.87***	
					(0.17)	
chinese:22					0.93***	
					(0.20)	
chinese:23					0.93***	
					(0.23)	
chinese:24					1.07***	
					(0.29)	
chinese:25					1.10**	
					(0.35)	
chinese:26					1.07*	
					(0.41)	
chinese:27					1.15*	
					(0.52)	
chinese:28					1.25	
					(0.74)	
chinese:29					1.21	
					(1.04)	
indo_iranian:1					0.33***	
					(0.07)	
indo_iranian:2					0.31***	
					(0.05)	
indo_iranian:3					0.32***	
					(0.04)	
indo_iranian:4					0.32***	
					(0.04)	
indo_iranian:5					0.26***	
					(0.03)	
indo_iranian:6					0.22***	
					(0.03)	
indo_iranian:7					0.16***	
					(0.03)	
indo_iranian:8					0.08**	
					(0.03)	
indo_iranian:9					0.01	
					(0.02)	
indo_iranian:10					−0.05*	
					(0.02)	
indo_iranian:11					−0.11***	
					(0.02)	
indo_iranian:12					−0.18***	
					(0.02)	
indo_iranian:13					−0.24***	
					(0.02)	
indo_iranian:14					−0.30***	
					(0.02)	
indo_iranian:15					−0.37***	
					(0.03)	
indo_iranian:16					−0.45***	
					(0.03)	
indo_iranian:17					−0.53***	
					(0.03)	

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
indo_iranian:18					-0.60^{***}	
					(0.03)	
indo_iranian:19					-0.67^{***}	
					(0.04)	
indo_iranian:20					-0.71^{***}	
					(0.04)	
indo_iranian:21					-0.78^{***}	
					(0.05)	
indo_iranian:22					-0.82^{***}	
					(0.06)	
indo_iranian:23					-0.84^{***}	
					(0.06)	
indo_iranian:24					-0.84^{***}	
					(0.08)	
indo_iranian:25					-0.85^{***}	
					(0.09)	
indo_iranian:26					-0.84^{***}	
					(0.11)	
indo_iranian:27					-0.84^{***}	
					(0.13)	
indo_iranian:28					-0.91^{***}	
					(0.17)	
indo_iranian:29					-0.97^{***}	
					(0.24)	
italic:1					-0.08	
					(0.06)	
italic:2					-0.13^{**}	
					(0.04)	
italic:3					-0.12^{***}	
					(0.04)	
italic:4					-0.13^{***}	
					(0.03)	
italic:5					-0.15^{***}	
					(0.03)	
italic:6					-0.15^{***}	
					(0.02)	
italic:7					-0.17^{***}	
					(0.02)	
italic:8					-0.20^{***}	
					(0.02)	
italic:9					-0.22^{***}	
					(0.02)	
italic:10					-0.23^{***}	
					(0.02)	
italic:11					-0.25^{***}	
					(0.02)	
italic:12					-0.28^{***}	
					(0.02)	
italic:13					-0.30^{***}	
					(0.02)	
italic:14					-0.32^{***}	
					(0.02)	
italic:15					-0.35^{***}	
					(0.02)	
italic:16					-0.38^{***}	
					(0.02)	
italic:17					-0.40^{***}	
					(0.03)	
italic:18					-0.42^{***}	
					(0.03)	
italic:19					-0.42^{***}	
					(0.03)	
italic:20					-0.41^{***}	
					(0.04)	
italic:21					-0.40^{***}	
					(0.04)	
italic:22					-0.38^{***}	
					(0.05)	
italic:23					-0.36^{***}	
					(0.06)	
italic:24					-0.31^{***}	
					(0.07)	
italic:25					-0.32^{***}	
					(0.08)	
italic:26					-0.29^{**}	

	m1__full	m1__base	m1__PCA	m1__sent	m1__lan	m1__TLD
					(0.10)	
italic:27					-0.28^{*}	
					(0.12)	
italic:28					-0.28	
					(0.17)	
italic:29					-0.30	
					(0.24)	
japanese:1					0.42^{***}	
					(0.13)	
japanese:2					0.37^{***}	
					(0.09)	
japanese:3					0.32^{***}	
					(0.08)	
japanese:4					0.27^{***}	
					(0.07)	
japanese:5					0.22^{***}	
					(0.06)	
japanese:6					0.14^{*}	
					(0.06)	
japanese:7					0.09	
					(0.05)	
japanese:8					0.04	
					(0.05)	
japanese:9					0.01	
					(0.05)	
japanese:10					-0.04	
					(0.05)	
japanese:11					-0.07	
					(0.05)	
japanese:12					-0.13^{**}	
					(0.05)	
japanese:13					-0.18^{***}	
					(0.05)	
japanese:14					-0.24^{***}	
					(0.05)	
japanese:15					-0.32^{***}	
					(0.05)	
japanese:16					-0.40^{***}	
					(0.05)	
japanese:17					-0.50^{***}	
					(0.06)	
japanese:18					-0.59^{***}	
					(0.06)	
japanese:19					-0.67^{***}	
					(0.07)	
japanese:20					-0.76^{***}	
					(0.07)	
japanese:21					-0.84^{***}	
					(0.08)	
japanese:22					-0.90^{***}	
					(0.09)	
japanese:23					-0.94^{***}	
					(0.10)	
japanese:24					-0.99^{***}	
					(0.11)	
japanese:25					-1.02^{***}	
					(0.13)	
japanese:26					-0.94^{***}	
					(0.16)	
japanese:27					-0.91^{***}	
					(0.19)	
japanese:28					-0.81^{**}	
					(0.27)	
japanese:29					-0.50	
					(0.43)	
other:1					-0.22^{*}	
					(0.11)	
other:2					-0.28^{***}	
					(0.08)	
other:3					-0.30^{***}	
					(0.07)	
other:4					-0.27^{***}	
					(0.06)	
other:5					-0.23^{***}	
					(0.05)	

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
other:6					-0.22^{***}	
					(0.04)	
other:7					-0.15^{***}	
					(0.04)	
other:8					-0.06	
					(0.03)	
other:9					0.06	
					(0.03)	
other:10					0.16^{***}	
					(0.03)	
other:11					0.24^{***}	
					(0.03)	
other:12					0.31^{***}	
					(0.03)	
other:13					0.40^{***}	
					(0.03)	
other:14					0.52^{***}	
					(0.04)	
other:15					0.61^{***}	
					(0.04)	
other:16					0.75^{***}	
					(0.05)	
other:17					0.86^{***}	
					(0.06)	
other:18					1.04^{***}	
					(0.08)	
other:19					1.20^{***}	
					(0.09)	
other:20					1.40^{***}	
					(0.12)	
other:21					1.58^{***}	
					(0.15)	
other:22					1.97^{***}	
					(0.22)	
other:23					2.26^{***}	
					(0.29)	
other:24					2.83^{***}	
					(0.45)	
other:25					2.99^{***}	
					(0.58)	
other:26					3.02^{***}	
					(0.71)	
other:27					15.08	
					(333.75)	
other:28					15.63	
					(595.07)	
other:29					15.87	
					(950.73)	
semitic:1					0.06	
					(0.08)	
semitic:2					0.06	
					(0.06)	
semitic:3					0.08	
					(0.05)	
semitic:4					0.08^{*}	
					(0.04)	
semitic:5					0.09^{*}	
					(0.03)	
semitic:6					0.10^{***}	
					(0.03)	
semitic:7					0.13^{***}	
					(0.03)	
semitic:8					0.16^{***}	
					(0.02)	
semitic:9					0.20^{***}	
					(0.02)	
semitic:10					0.23^{***}	
					(0.02)	
semitic:11					0.27^{***}	
					(0.02)	
semitic:12					0.30^{***}	
					(0.02)	
semitic:13					0.35^{***}	
					(0.03)	
semitic:14					0.41^{***}	

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
					(0.03)	
semitic:15					0.49***	
					(0.03)	
semitic:16					0.58***	
					(0.04)	
semitic:17					0.66***	
					(0.05)	
semitic:18					0.79***	
					(0.06)	
semitic:19					0.99***	
					(0.07)	
semitic:20					1.14***	
					(0.09)	
semitic:21					1.29***	
					(0.11)	
semitic:22					1.56***	
					(0.15)	
semitic:23					1.57***	
					(0.17)	
semitic:24					1.85***	
					(0.23)	
semitic:25					2.20***	
					(0.33)	
semitic:26					2.55***	
					(0.47)	
semitic:27					2.63***	
					(0.58)	
semitic:28					2.36***	
					(0.71)	
semitic:29					2.35*	
					(1.00)	
slavic:1					0.20**	
					(0.07)	
slavic:2					0.04	
					(0.05)	
slavic:3					−0.01	
					(0.05)	
slavic:4					−0.09*	
					(0.04)	
slavic:5					−0.10**	
					(0.03)	
slavic:6					−0.14***	
					(0.03)	
slavic:7					−0.14***	
					(0.03)	
slavic:8					−0.14***	
					(0.02)	
slavic:9					−0.12***	
					(0.02)	
slavic:10					−0.11***	
					(0.02)	
slavic:11					−0.09***	
					(0.02)	
slavic:12					−0.09***	
					(0.02)	
slavic:13					−0.08***	
					(0.02)	
slavic:14					−0.07**	
					(0.02)	
slavic:15					−0.05	
					(0.03)	
slavic:16					−0.01	
					(0.03)	
slavic:17					0.05	
					(0.03)	
slavic:18					0.11**	
					(0.04)	
slavic:19					0.21***	
					(0.05)	
slavic:20					0.38***	
					(0.06)	
slavic:21					0.56***	
					(0.07)	
slavic:22					0.67***	
					(0.09)	

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
slavic:23					0.86***	
					(0.11)	
slavic:24					1.03***	
					(0.14)	
slavic:25					1.26***	
					(0.19)	
slavic:26					1.46***	
					(0.25)	
slavic:27					1.81***	
					(0.35)	
slavic:28					2.24***	
					(0.59)	
slavic:29					3.70*	
					(1.63)	
turkic:1					0.21	
					(0.13)	
turkic:2					0.13	
					(0.10)	
turkic:3					0.07	
					(0.09)	
turkic:4					0.07	
					(0.07)	
turkic:5					0.08	
					(0.06)	
turkic:6					0.09	
					(0.06)	
turkic:7					0.10	
					(0.05)	
turkic:8					0.10*	
					(0.05)	
turkic:9					0.10*	
					(0.04)	
turkic:10					0.10*	
					(0.04)	
turkic:11					0.12**	
					(0.04)	
turkic:12					0.14***	
					(0.04)	
turkic:13					0.18***	
					(0.05)	
turkic:14					0.23***	
					(0.05)	
turkic:15					0.28***	
					(0.06)	
turkic:16					0.34***	
					(0.06)	
turkic:17					0.42***	
					(0.07)	
turkic:18					0.51***	
					(0.09)	
turkic:19					0.65***	
					(0.11)	
turkic:20					0.87***	
					(0.14)	
turkic:21					1.04***	
					(0.17)	
turkic:22					1.45***	
					(0.24)	
turkic:23					1.42***	
					(0.28)	
turkic:24					1.32***	
					(0.31)	
turkic:25					1.15***	
					(0.35)	
turkic:26					1.11**	
					(0.41)	
turkic:27					1.23*	
					(0.51)	
turkic:28					1.32	
					(0.72)	
turkic:29					1.28	
					(1.01)	
TLD_cz:1						−0.20
						(0.36)
TLD_cz:2						−0.27***

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
						(0.02)
TLD_cz:3						-0.31^{***}
						(0.02)
TLD_cz:4						-0.31^{**}
						(0.10)
TLD_cz:5						-0.31^{*}
						(0.13)
TLD_cz:6						-0.23
						(0.14)
TLD_cz:7						-0.02
						(0.13)
TLD_cz:8						0.05
						(0.12)
TLD_cz:9						0.16
						(0.11)
TLD_cz:10						0.28^{**}
						(0.10)
TLD_cz:11						0.41^{***}
						(0.10)
TLD_cz:12						0.49^{***}
						(0.10)
TLD_cz:13						0.67^{***}
						(0.11)
TLD_cz:14						0.78^{***}
						(0.12)
TLD_cz:15						0.97^{***}
						(0.14)
TLD_cz:16						1.23^{***}
						(0.17)
TLD_cz:17						1.56^{***}
						(0.21)
TLD_cz:18						1.72^{***}
						(0.27)
TLD_cz:19						1.72^{***}
						(0.31)
TLD_cz:20						1.72^{***}
						(0.42)
TLD_cz:21						2.01^{***}
						(0.51)
TLD_cz:22						2.01^{*}
						(0.81)
TLD_cz:23						2.61
						(1.91)
TLD_cz:24						3.81
						(4.19)
TLD_es:1						0.11
						(0.24)
TLD_es:2						0.11
						(0.12)
TLD_es:3						0.18
						(0.11)
TLD_es:4						0.19
						(0.10)
TLD_es:5						0.19
						(0.10)
TLD_es:6						0.20^{*}
						(0.09)
TLD_es:7						0.21^{*}
						(0.08)
TLD_es:8						0.21^{**}
						(0.08)
TLD_es:9						0.19^{**}
						(0.07)
TLD_es:10						0.18^{**}
						(0.07)
TLD_es:11						0.15^{*}
						(0.07)
TLD_es:12						0.11
						(0.07)
TLD_es:13						0.10
						(0.07)
TLD_es:14						0.07
						(0.07)
TLD_es:15						0.04
						(0.07)

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
TLD_es:16						−0.01
						(0.08)
TLD_es:17						−0.07
						(0.09)
TLD_es:18						−0.17
						(0.10)
TLD_es:19						−0.27*
						(0.11)
TLD_es:20						−0.42**
						(0.14)
TLD_es:21						−0.44**
						(0.15)
TLD_es:22						−0.46*
						(0.21)
TLD_es:23						−0.58
						(0.34)
TLD_es:24						−0.58
						(0.42)
TLD_fr:1						0.47*
						(0.20)
TLD_fr:2						0.44***
						(0.02)
TLD_fr:3						0.42***
						(0.02)
TLD_fr:4						0.24***
						(0.05)
TLD_fr:5						0.18**
						(0.07)
TLD_fr:6						0.10
						(0.07)
TLD_fr:7						0.10
						(0.08)
TLD_fr:8						0.06
						(0.07)
TLD_fr:9						0.03
						(0.07)
TLD_fr:10						0.02
						(0.07)
TLD_fr:11						0.02
						(0.07)
TLD_fr:12						0.01
						(0.07)
TLD_fr:13						0.01
						(0.07)
TLD_fr:14						0.01
						(0.07)
TLD_fr:15						0.01
						(0.08)
TLD_fr:16						0.01
						(0.08)
TLD_fr:17						−0.00
						(0.09)
TLD_fr:18						−0.02
						(0.11)
TLD_fr:19						−0.02
						(0.12)
TLD_fr:20						−0.02
						(0.16)
TLD_fr:21						−0.02
						(0.17)
TLD_fr:22						−0.02
						(0.24)
TLD_fr:23						0.08
						(0.40)
TLD_fr:24						0.08
						(0.50)
TLD_hu:1						0.42
						(0.27)
TLD_hu:2						0.39***
						(0.02)
TLD_hu:3						0.33***
						(0.02)
TLD_hu:4						0.33***
						(0.09)
TLD_hu:5						0.33**

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
						(0.10)
TLD_hu:6						0.33**
						(0.11)
TLD_hu:7						0.37***
						(0.10)
TLD_hu:8						0.44***
						(0.10)
TLD_hu:9						0.52***
						(0.09)
TLD_hu:10						0.59***
						(0.09)
TLD_hu:11						0.69***
						(0.09)
TLD_hu:12						0.87***
						(0.10)
TLD_hu:13						1.06***
						(0.11)
TLD_hu:14						1.29***
						(0.12)
TLD_hu:15						1.60***
						(0.16)
TLD_hu:16						2.07***
						(0.22)
TLD_hu:17						2.42***
						(0.30)
TLD_hu:18						2.72***
						(0.42)
TLD_hu:19						3.25***
						(0.61)
TLD_hu:20						3.33***
						(0.86)
TLD_hu:21						3.42***
						(0.96)
TLD_hu:22						14.50
						(445.50)
TLD_hu:23						15.35
						(938.90)
TLD_hu:24						15.59
						(1307.05)
TLD_it:1						0.58**
						(0.19)
TLD_it:2						0.54***
						(0.02)
TLD_it:3						0.49***
						(0.02)
TLD_it:4						0.46***
						(0.05)
TLD_it:5						0.44***
						(0.07)
TLD_it:6						0.33***
						(0.07)
TLD_it:7						0.32***
						(0.07)
TLD_it:8						0.27***
						(0.07)
TLD_it:9						0.23***
						(0.07)
TLD_it:10						0.19**
						(0.06)
TLD_it:11						0.15*
						(0.06)
TLD_it:12						0.11
						(0.06)
TLD_it:13						0.11
						(0.06)
TLD_it:14						0.11
						(0.07)
TLD_it:15						0.12
						(0.07)
TLD_it:16						0.15
						(0.08)
TLD_it:17						0.18*
						(0.09)
TLD_it:18						0.24*
						(0.11)

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
TLD_it:19						0.33**
						(0.12)
TLD_it:20						0.35*
						(0.16)
TLD_it:21						0.52**
						(0.18)
TLD_it:22						0.52*
						(0.26)
TLD_it:23						0.52
						(0.43)
TLD_it:24						0.52
						(0.53)
TLD_other:1						0.13
						(0.17)
TLD_other:2						0.13***
						(0.02)
TLD_other:3						0.13***
						(0.02)
TLD_other:4						0.13**
						(0.05)
TLD_other:5						0.14*
						(0.06)
TLD_other:6						0.15*
						(0.06)
TLD_other:7						0.24***
						(0.06)
TLD_other:8						0.28***
						(0.06)
TLD_other:9						0.33***
						(0.05)
TLD_other:10						0.36***
						(0.05)
TLD_other:11						0.39***
						(0.05)
TLD_other:12						0.44***
						(0.05)
TLD_other:13						0.52***
						(0.05)
TLD_other:14						0.59***
						(0.05)
TLD_other:15						0.66***
						(0.06)
TLD_other:16						0.75***
						(0.06)
TLD_other:17						0.83***
						(0.07)
TLD_other:18						0.91***
						(0.08)
TLD_other:19						0.99***
						(0.09)
TLD_other:20						1.07***
						(0.13)
TLD_other:21						1.15***
						(0.14)
TLD_other:22						1.15***
						(0.20)
TLD_other:23						1.15***
						(0.33)
TLD_other:24						1.15**
						(0.40)
TLD_pl:1						0.49*
						(0.24)
TLD_pl:2						0.49***
						(0.11)
TLD_pl:3						0.49***
						(0.11)
TLD_pl:4						0.49***
						(0.11)
TLD_pl:5						0.48***
						(0.10)
TLD_pl:6						0.43***
						(0.10)
TLD_pl:7						0.43***
						(0.09)
TLD_pl:8						0.41***

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
						(0.08)
TLD_pl:9						0.41^{***}
						(0.08)
TLD_pl:10						0.41^{***}
						(0.07)
TLD_pl:11						0.41^{***}
						(0.07)
TLD_pl:12						0.41^{***}
						(0.07)
TLD_pl:13						0.43^{***}
						(0.08)
TLD_pl:14						0.44^{***}
						(0.08)
TLD_pl:15						0.45^{***}
						(0.09)
TLD_pl:16						0.48^{***}
						(0.10)
TLD_pl:17						0.57^{***}
						(0.11)
TLD_pl:18						0.58^{***}
						(0.13)
TLD_pl:19						0.60^{***}
						(0.15)
TLD_pl:20						0.68^{***}
						(0.20)
TLD_pl:21						0.82^{***}
						(0.23)
TLD_pl:22						0.82^{*}
						(0.34)
TLD_pl:23						0.91
						(0.57)
TLD_pl:24						1.03
						(0.74)
TLD_se:1						-0.32
						(0.34)
TLD_se:2						-0.32
						(0.18)
TLD_se:3						-0.32
						(0.18)
TLD_se:4						-0.32
						(0.16)
TLD_se:5						-0.18
						(0.14)
TLD_se:6						-0.10
						(0.12)
TLD_se:7						0.06
						(0.10)
TLD_se:8						0.12
						(0.09)
TLD_se:9						0.13
						(0.08)
TLD_se:10						0.13
						(0.08)
TLD_se:11						0.17^{*}
						(0.08)
TLD_se:12						0.19^{*}
						(0.08)
TLD_se:13						0.26^{***}
						(0.08)
TLD_se:14						0.32^{***}
						(0.08)
TLD_se:15						0.38^{***}
						(0.09)
TLD_se:16						0.46^{***}
						(0.10)
TLD_se:17						0.55^{***}
						(0.12)
TLD_se:18						0.66^{***}
						(0.14)
TLD_se:19						0.71^{***}
						(0.16)
TLD_se:20						0.76^{***}
						(0.22)
TLD_se:21						0.77^{**}
						(0.23)

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
TLD_se:22						0.77*
						(0.35)
TLD_se:23						0.77
						(0.58)
TLD_se:24						0.77
						(0.71)
TLD_sk:1						0.53
						(0.27)
TLD_sk:2						0.47***
						(0.02)
TLD_sk:3						0.44***
						(0.02)
TLD_sk:4						0.27***
						(0.05)
TLD_sk:5						0.25**
						(0.09)
TLD_sk:6						0.25*
						(0.11)
TLD_sk:7						0.31**
						(0.11)
TLD_sk:8						0.40***
						(0.10)
TLD_sk:9						0.51***
						(0.10)
TLD_sk:10						0.58***
						(0.09)
TLD_sk:11						0.72***
						(0.10)
TLD_sk:12						0.79***
						(0.10)
TLD_sk:13						0.92***
						(0.11)
TLD_sk:14						1.09***
						(0.12)
TLD_sk:15						1.25***
						(0.15)
TLD_sk:16						1.47***
						(0.18)
TLD_sk:17						1.75***
						(0.23)
TLD_sk:18						2.14***
						(0.33)
TLD_sk:19						2.15***
						(0.38)
TLD_sk:20						2.24***
						(0.53)
TLD_sk:21						2.55***
						(0.66)
TLD_sk:22						3.20*
						(1.40)
TLD_sk:23						15.24
						(976.37)
TLD_sk:24						15.51
						(1380.75)
TLD_uk:1						0.60**
						(0.19)
TLD_uk:2						0.60***
						(0.02)
TLD_uk:3						0.56***
						(0.02)
TLD_uk:4						0.52***
						(0.05)
TLD_uk:5						0.51***
						(0.07)
TLD_uk:6						0.43***
						(0.07)
TLD_uk:7						0.43***
						(0.07)
TLD_uk:8						0.39***
						(0.07)
TLD_uk:9						0.34***
						(0.06)
TLD_uk:10						0.31***
						(0.06)
TLD_uk:11						0.28***

	m1_full	m1_base	m1_PCA	m1_sent	m1_lan	m1_TLD
						(0.06)
TLD_uk:12						0.27***
						(0.06)
TLD_uk:13						0.26***
						(0.06)
TLD_uk:14						0.26***
						(0.07)
TLD_uk:15						0.24**
						(0.07)
TLD_uk:16						0.21**
						(0.08)
TLD_uk:17						0.17
						(0.09)
TLD_uk:18						0.12
						(0.10)
TLD_uk:19						0.09
						(0.11)
TLD_uk:20						0.08
						(0.15)
TLD_uk:21						0.06
						(0.16)
TLD_uk:22						0.02
						(0.23)
TLD_uk:23						0.02
						(0.38)
TLD_uk:24						0.02
						(0.47)
Log Likelihood	−499842.38	−342622.84	−353591.26	−113277.37	−347533.40	
DF	2117783	1310684	1372193	481313	1346557	258120
Num. obs.	2118015	1310858	1372454	481545	1346992	258504

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A.10: Statistical models

A.5 Model family 2 estimates

	m2_full	m2_base	m2_lan	m2_TLD
(Intercept):1	−2.50***	−2.53***	−2.61***	−2.60***
	(0.22)	(0.16)	(0.21)	(0.42)
(Intercept):2	−1.81***	−1.79***	−1.77***	−1.84***
	(0.16)	(0.12)	(0.15)	(0.29)
(Intercept):3	−1.43***	−1.39***	−1.36***	−1.45***
	(0.14)	(0.10)	(0.13)	(0.25)
(Intercept):4	−1.12***	−1.09***	−1.05***	−1.16***
	(0.12)	(0.09)	(0.12)	(0.22)
(Intercept):5	−0.87***	−0.81***	−0.77***	−0.89***
	(0.12)	(0.09)	(0.11)	(0.21)
(Intercept):6	−0.62***	−0.56***	−0.53***	−0.67***
	(0.11)	(0.08)	(0.10)	(0.20)
(Intercept):7	−0.29**	−0.24**	−0.20*	−0.36
	(0.11)	(0.08)	(0.10)	(0.19)

	m2_full	m2_base	m2_lan	m2_TLD
(Intercept):8	0.08	0.10	0.15	-0.03
	(0.11)	(0.08)	(0.10)	(0.19)
(Intercept):9	0.52***	0.50***	0.55***	0.36
	(0.11)	(0.08)	(0.10)	(0.19)
(Intercept):10	0.89***	0.84***	0.90***	0.69***
	(0.11)	(0.08)	(0.10)	(0.19)
(Intercept):11	1.27***	1.20***	1.26***	1.04***
	(0.11)	(0.09)	(0.11)	(0.19)
(Intercept):12	1.55***	1.47***	1.54***	1.30***
	(0.12)	(0.09)	(0.12)	(0.20)
(Intercept):13	1.82***	1.74***	1.81***	1.56***
	(0.13)	(0.10)	(0.13)	(0.21)
(Intercept):14	2.06***	1.97***	2.04***	1.78***
	(0.14)	(0.11)	(0.14)	(0.23)
(Intercept):15	2.35***	2.30***	2.38***	2.14***
	(0.15)	(0.12)	(0.15)	(0.25)
(Intercept):16	3.16***	2.99***	3.10***	2.89***
	(0.19)	(0.15)	(0.19)	(0.29)
(Intercept):17	3.62***	3.48***	3.66***	3.58***
	(0.21)	(0.17)	(0.22)	(0.35)
(Intercept):18	3.85***	3.74***	3.94***	4.00***
	(0.23)	(0.19)	(0.25)	(0.43)
(Intercept):19	4.15***	4.06***	4.34***	4.82***
	(0.27)	(0.22)	(0.29)	(0.63)
(Intercept):20	4.51***	4.41***	4.81***	
	(0.30)	(0.25)	(0.35)	
(Intercept):21	4.99***	4.88***	5.55***	
	(0.35)	(0.29)	(0.45)	
(Intercept):22	5.39***	5.25***	6.18***	
	(0.42)	(0.34)	(0.57)	
(Intercept):23	5.62***	5.49***	6.60***	
	(0.47)	(0.39)	(0.73)	
(Intercept):24	5.70***	5.67***	7.21***	
	(0.52)	(0.42)	(0.90)	
(Intercept):25	5.67***	5.72***		
	(0.59)	(0.45)		
(Intercept):26	5.65***	5.72***		
	(0.60)	(0.45)		

	m2_full	m2_base	m2_lan	m2_TLD
(Intercept):27	5.67***	5.94***		
	(0.66)	(0.51)		
(Intercept):28	5.89***	6.26***		
	(0.79)	(0.62)		
(Intercept):29	6.39***	7.14***		
	(1.12)	(0.96)		
Cyber:1	−0.39	−0.50*	−0.47	−0.40
	(0.21)	(0.24)	(0.27)	(0.21)
Cyber:2	−0.35*	−0.41*	−0.42*	−0.31*
	(0.15)	(0.18)	(0.19)	(0.15)
Cyber:3	−0.26*	−0.33*	−0.33*	−0.22
	(0.13)	(0.15)	(0.17)	(0.13)
Cyber:4	−0.20	−0.30*	−0.29	−0.18
	(0.12)	(0.14)	(0.15)	(0.11)
Cyber:5	−0.20	−0.30*	−0.29*	−0.16
	(0.11)	(0.13)	(0.14)	(0.11)
Cyber:6	−0.20	−0.29*	−0.28*	−0.12
	(0.11)	(0.12)	(0.13)	(0.10)
Cyber:7	−0.28**	−0.36**	−0.35**	−0.20*
	(0.10)	(0.12)	(0.13)	(0.10)
Cyber:8	−0.40***	−0.45***	−0.45***	−0.29**
	(0.10)	(0.12)	(0.13)	(0.10)
Cyber:9	−0.57***	−0.60***	−0.61***	−0.43***
	(0.10)	(0.12)	(0.13)	(0.10)
Cyber:10	−0.71***	−0.71***	−0.74***	−0.54***
	(0.10)	(0.12)	(0.13)	(0.10)
Cyber:11	−0.86***	−0.84***	−0.89***	−0.66***
	(0.11)	(0.13)	(0.14)	(0.11)
Cyber:12	−0.90***	−0.90***	−0.95***	−0.69***
	(0.12)	(0.13)	(0.15)	(0.12)
Cyber:13	−0.94***	−0.97***	−1.01***	−0.72***
	(0.12)	(0.14)	(0.16)	(0.13)
Cyber:14	−0.95***	−0.99***	−1.03***	−0.69***
	(0.13)	(0.15)	(0.17)	(0.14)
Cyber:15	−1.04***	−1.13***	−1.16***	−0.78***
	(0.14)	(0.17)	(0.19)	(0.15)
Cyber:16	−1.46***	−1.57***	−1.65***	−1.26***
	(0.18)	(0.20)	(0.23)	(0.20)

	m2_full	m2_base	m2_lan	m2_TLD
Cyber:17	-1.70^{***}	-1.86^{***}	-2.00^{***}	-1.59^{***}
	(0.20)	(0.23)	(0.27)	(0.25)
Cyber:18	-1.75^{***}	-1.92^{***}	-2.05^{***}	-1.53^{***}
	(0.22)	(0.26)	(0.30)	(0.32)
Cyber:19	-1.88^{***}	-2.05^{***}	-2.20^{***}	-1.56^{***}
	(0.25)	(0.29)	(0.35)	(0.47)
Cyber:20	-2.08^{***}	-2.21^{***}	-2.40^{***}	
	(0.29)	(0.33)	(0.42)	
Cyber:21	-2.38^{***}	-2.48^{***}	-2.85^{***}	
	(0.34)	(0.38)	(0.53)	
Cyber:22	-2.49^{***}	-2.60^{***}	-3.15^{***}	
	(0.40)	(0.45)	(0.67)	
Cyber:23	-2.55^{***}	-2.61^{***}	-3.07^{***}	
	(0.46)	(0.50)	(0.87)	
Cyber:24	-2.52^{***}	-2.57^{***}	-2.92^{**}	
	(0.50)	(0.55)	(1.05)	
Cyber:25	-2.53^{***}	-2.42^{***}		
	(0.60)	(0.58)		
Cyber:26	-2.23^{***}	-2.08^{***}		
	(0.64)	(0.59)		
Cyber:27	-2.12^{**}	-2.01^{**}		
	(0.71)	(0.68)		
Cyber:28	-2.01^{*}	-1.92^{*}		
	(0.83)	(0.82)		
Cyber:29	-2.13	-2.11		
	(1.18)	(1.26)		
Mobile:1	-0.06			
	(0.18)			
Mobile:2	-0.01			
	(0.13)			
Mobile:3	0.01			
	(0.11)			
Mobile:4	0.02			
	(0.10)			
Mobile:5	0.03			
	(0.09)			
Mobile:6	0.03			
	(0.09)			

	m2_full	m2_base	m2_lan	m2_TLD
Mobile:7	0.03			
	(0.09)			
Mobile:8	0.03			
	(0.09)			
Mobile:9	0.01			
	(0.09)			
Mobile:10	0.00			
	(0.09)			
Mobile:11	−0.01			
	(0.09)			
Mobile:12	−0.02			
	(0.10)			
Mobile:13	−0.02			
	(0.10)			
Mobile:14	−0.03			
	(0.11)			
Mobile:15	−0.03			
	(0.12)			
Mobile:16	−0.08			
	(0.14)			
Mobile:17	−0.10			
	(0.16)			
Mobile:18	−0.10			
	(0.17)			
Mobile:19	−0.06			
	(0.19)			
Mobile:20	−0.02			
	(0.22)			
Mobile:21	−0.01			
	(0.25)			
Mobile:22	0.01			
	(0.29)			
Mobile:23	0.07			
	(0.33)			
Mobile:24	0.18			
	(0.36)			
Mobile:25	0.53			
	(0.44)			

	m2_full	m2_base	m2_lan	m2_TLD
Mobile:26	0.58			
	(0.47)			
Mobile:27	0.70			
	(0.53)			
Mobile:28	0.73			
	(0.64)			
Mobile:29	0.93			
	(0.90)			
SexF:1	−0.03			
	(0.09)			
SexF:2	−0.02			
	(0.07)			
SexF:3	−0.02			
	(0.06)			
SexF:4	−0.03			
	(0.05)			
SexF:5	−0.03			
	(0.05)			
SexF:6	−0.03			
	(0.05)			
SexF:7	−0.01			
	(0.04)			
SexF:8	0.01			
	(0.04)			
SexF:9	0.04			
	(0.04)			
SexF:10	0.06			
	(0.04)			
SexF:11	0.07			
	(0.05)			
SexF:12	0.07			
	(0.05)			
SexF:13	0.07			
	(0.05)			
SexF:14	0.06			
	(0.05)			
SexF:15	0.06			
	(0.06)			

	m2_full	m2_base	m2_lan	m2_TLD
SexF:16	0.06			
	(0.07)			
SexF:17	0.04			
	(0.08)			
SexF:18	0.04			
	(0.08)			
SexF:19	0.03			
	(0.09)			
SexF:20	0.03			
	(0.10)			
SexF:21	−0.00			
	(0.12)			
SexF:22	−0.05			
	(0.13)			
SexF:23	−0.05			
	(0.15)			
SexF:24	−0.07			
	(0.17)			
SexF:25	−0.08			
	(0.20)			
SexF:26	−0.09			
	(0.22)			
SexF:27	−0.09			
	(0.25)			
SexF:28	−0.08			
	(0.30)			
SexF:29	−0.11			
	(0.42)			
austro_asiatic:1			0.10	
			(0.53)	
austro_asiatic:2			0.04	
			(0.39)	
austro_asiatic:3			0.02	
			(0.36)	
austro_asiatic:4			−0.01	
			(0.34)	
austro_asiatic:5			−0.03	
			(0.32)	

	m2_full	m2_base	m2_lan	m2_TLD
austro_asiatic:6			−0.06	
			(0.31)	
austro_asiatic:7			−0.10	
			(0.29)	
austro_asiatic:8			−0.15	
			(0.26)	
austro_asiatic:9			−0.22	
			(0.26)	
austro_asiatic:10			−0.24	
			(0.26)	
austro_asiatic:11			−0.28	
			(0.26)	
austro_asiatic:12			−0.30	
			(0.26)	
austro_asiatic:13			−0.32	
			(0.28)	
austro_asiatic:14			−0.33	
			(0.30)	
austro_asiatic:15			−0.36	
			(0.30)	
austro_asiatic:16			−0.48	
			(0.30)	
austro_asiatic:17			−0.52	
			(0.34)	
austro_asiatic:18			−0.51	
			(0.37)	
austro_asiatic:19			−0.56	
			(0.42)	
austro_asiatic:20			−0.59	
			(0.48)	
austro_asiatic:21			−0.66	
			(0.54)	
austro_asiatic:22			−0.66	
			(0.70)	
austro_asiatic:23			−0.87	
			(0.73)	
austro_asiatic:24			−0.85	
			(1.13)	

	m2_full	m2_base	m2_lan	m2_TLD
chinese:1			0.03	
			(0.58)	
chinese:2			0.01	
			(0.40)	
chinese:3			0.00	
			(0.36)	
chinese:4			0.01	
			(0.30)	
chinese:5			0.01	
			(0.28)	
chinese:6			0.02	
			(0.26)	
chinese:7			0.01	
			(0.26)	
chinese:8			−0.00	
			(0.25)	
chinese:9			−0.02	
			(0.25)	
chinese:10			−0.02	
			(0.24)	
chinese:11			−0.01	
			(0.24)	
chinese:12			0.01	
			(0.24)	
chinese:13			0.05	
			(0.25)	
chinese:14			0.09	
			(0.27)	
chinese:15			0.14	
			(0.30)	
chinese:16			0.13	
			(0.33)	
chinese:17			0.18	
			(0.40)	
chinese:18			0.26	
			(0.46)	
chinese:19			0.27	
			(0.52)	

	m2_full	m2_base	m2_lan	m2_TLD
chinese:20			0.44	
			(0.65)	
chinese:21			0.66	
			(0.82)	
chinese:22			1.27	
			(1.34)	
chinese:23			0.94	
			(1.45)	
chinese:24			0.28	
			(1.49)	
indo_iranian:1			−0.03	
			(0.29)	
indo_iranian:2			−0.08	
			(0.21)	
indo_iranian:3			−0.07	
			(0.18)	
indo_iranian:4			−0.07	
			(0.16)	
indo_iranian:5			−0.07	
			(0.15)	
indo_iranian:6			−0.06	
			(0.14)	
indo_iranian:7			−0.07	
			(0.14)	
indo_iranian:8			−0.09	
			(0.13)	
indo_iranian:9			−0.12	
			(0.13)	
indo_iranian:10			−0.15	
			(0.13)	
indo_iranian:11			−0.17	
			(0.14)	
indo_iranian:12			−0.18	
			(0.14)	
indo_iranian:13			−0.19	
			(0.15)	
indo_iranian:14			−0.19	
			(0.15)	

	m2_full	m2_base	m2_lan	m2_TLD
indo_iranian:15			−0.21	
			(0.16)	
indo_iranian:16			−0.28	
			(0.18)	
indo_iranian:17			−0.35	
			(0.20)	
indo_iranian:18			−0.38	
			(0.22)	
indo_iranian:19			−0.47	
			(0.24)	
indo_iranian:20			−0.55*	
			(0.26)	
indo_iranian:21			−0.62*	
			(0.31)	
indo_iranian:22			−0.64	
			(0.37)	
indo_iranian:23			−0.69	
			(0.46)	
indo_iranian:24			−0.62	
			(0.69)	
italic:1			0.00	
			(0.17)	
italic:2			−0.05	
			(0.12)	
italic:3			−0.05	
			(0.10)	
italic:4			−0.06	
			(0.09)	
italic:5			−0.06	
			(0.09)	
italic:6			−0.06	
			(0.08)	
italic:7			−0.07	
			(0.08)	
italic:8			−0.08	
			(0.08)	
italic:9			−0.10	
			(0.08)	

	m2__full	m2__base	m2__lan	m2__TLD
italic:10			−0.12	
			(0.08)	
italic:11			−0.13	
			(0.08)	
italic:12			−0.14	
			(0.09)	
italic:13			−0.14	
			(0.09)	
italic:14			−0.13	
			(0.10)	
italic:15			−0.13	
			(0.11)	
italic:16			−0.17	
			(0.12)	
italic:17			−0.22	
			(0.14)	
italic:18			−0.21	
			(0.15)	
italic:19			−0.21	
			(0.17)	
italic:20			−0.22	
			(0.20)	
italic:21			−0.24	
			(0.23)	
italic:22			−0.21	
			(0.29)	
italic:23			0.01	
			(0.39)	
italic:24			0.08	
			(0.57)	
japanese:1			−0.25	
			(0.70)	
japanese:2			−0.15	
			(0.42)	
japanese:3			−0.12	
			(0.34)	
japanese:4			−0.11	
			(0.32)	

	m2_full	m2_base	m2_lan	m2_TLD
japanese:5			−0.08	
			(0.29)	
japanese:6			−0.08	
			(0.27)	
japanese:7			−0.08	
			(0.24)	
japanese:8			−0.10	
			(0.24)	
japanese:9			−0.11	
			(0.23)	
japanese:10			−0.11	
			(0.24)	
japanese:11			−0.12	
			(0.25)	
japanese:12			−0.12	
			(0.26)	
japanese:13			−0.12	
			(0.26)	
japanese:14			−0.12	
			(0.28)	
japanese:15			−0.12	
			(0.30)	
japanese:16			−0.13	
			(0.34)	
japanese:17			−0.11	
			(0.38)	
japanese:18			−0.11	
			(0.41)	
japanese:19			−0.10	
			(0.47)	
japanese:20			−0.10	
			(0.54)	
japanese:21			−0.08	
			(0.65)	
japanese:22			−0.17	
			(0.77)	
japanese:23			−0.21	
			(1.03)	

	m2_full	m2_base	m2_lan	m2_TLD
japanese:24			−0.40	
			(1.49)	
other:1			0.01	
			(0.25)	
other:2			0.08	
			(0.17)	
other:3			0.09	
			(0.14)	
other:4			0.08	
			(0.13)	
other:5			0.09	
			(0.12)	
other:6			0.11	
			(0.11)	
other:7			0.13	
			(0.11)	
other:8			0.17	
			(0.11)	
other:9			0.23*	
			(0.11)	
other:10			0.30**	
			(0.11)	
other:11			0.37**	
			(0.12)	
other:12			0.42**	
			(0.13)	
other:13			0.46***	
			(0.14)	
other:14			0.53***	
			(0.15)	
other:15			0.62***	
			(0.17)	
other:16			0.79***	
			(0.21)	
other:17			0.96***	
			(0.25)	
other:18			1.24***	
			(0.31)	

	m2_full	m2_base	m2_lan	m2_TLD
other:19			1.53***	
			(0.39)	
other:20			1.67***	
			(0.50)	
other:21			1.86**	
			(0.66)	
other:22			1.97*	
			(0.86)	
other:23			1.91	
			(1.09)	
other:24			1.27	
			(1.16)	
semitic:1			0.16	
			(0.22)	
semitic:2			0.10	
			(0.16)	
semitic:3			0.06	
			(0.14)	
semitic:4			0.06	
			(0.12)	
semitic:5			0.07	
			(0.11)	
semitic:6			0.11	
			(0.11)	
semitic:7			0.13	
			(0.11)	
semitic:8			0.15	
			(0.10)	
semitic:9			0.20	
			(0.11)	
semitic:10			0.23*	
			(0.11)	
semitic:11			0.27*	
			(0.12)	
semitic:12			0.28*	
			(0.12)	
semitic:13			0.31*	
			(0.13)	

	m2__full	m2__base	m2__lan	m2__TLD
semitic:14			0.33*	
			(0.14)	
semitic:15			0.40*	
			(0.16)	
semitic:16			0.58**	
			(0.20)	
semitic:17			0.80**	
			(0.24)	
semitic:18			0.84**	
			(0.28)	
semitic:19			0.99**	
			(0.34)	
semitic:20			1.15**	
			(0.42)	
semitic:21			1.78**	
			(0.69)	
semitic:22			1.73*	
			(0.82)	
semitic:23			2.32	
			(1.36)	
semitic:24			4.22	
			(5.08)	
slavic:1			−0.01	
			(0.21)	
slavic:2			−0.03	
			(0.15)	
slavic:3			−0.04	
			(0.12)	
slavic:4			−0.05	
			(0.11)	
slavic:5			−0.06	
			(0.10)	
slavic:6			−0.06	
			(0.10)	
slavic:7			−0.07	
			(0.09)	
slavic:8			−0.08	
			(0.09)	

	m2_full	m2_base	m2_lan	m2_TLD
slavic:9			−0.09	
			(0.09)	
slavic:10			−0.11	
			(0.09)	
slavic:11			−0.12	
			(0.10)	
slavic:12			−0.13	
			(0.10)	
slavic:13			−0.13	
			(0.10)	
slavic:14			−0.13	
			(0.11)	
slavic:15			−0.13	
			(0.12)	
slavic:16			−0.20	
			(0.13)	
slavic:17			−0.21	
			(0.15)	
slavic:18			−0.22	
			(0.16)	
slavic:19			−0.22	
			(0.18)	
slavic:20			−0.24	
			(0.21)	
slavic:21			−0.29	
			(0.25)	
slavic:22			−0.46	
			(0.28)	
slavic:23			−0.51	
			(0.35)	
slavic:24			−0.61	
			(0.50)	
turkic:1			0.21	
			(0.49)	
turkic:2			0.16	
			(0.36)	
turkic:3			0.05	
			(0.33)	

	m2_full	m2_base	m2_lan	m2_TLD
turkic:4			−0.03	
			(0.31)	
turkic:5			0.01	
			(0.24)	
turkic:6			0.00	
			(0.23)	
turkic:7			0.03	
			(0.22)	
turkic:8			0.04	
			(0.21)	
turkic:9			0.07	
			(0.21)	
turkic:10			0.11	
			(0.22)	
turkic:11			0.13	
			(0.23)	
turkic:12			0.11	
			(0.24)	
turkic:13			0.15	
			(0.26)	
turkic:14			0.19	
			(0.27)	
turkic:15			0.24	
			(0.30)	
turkic:16			0.28	
			(0.35)	
turkic:17			0.34	
			(0.41)	
turkic:18			0.39	
			(0.47)	
turkic:19			0.49	
			(0.57)	
turkic:20			0.51	
			(0.69)	
turkic:21			0.82	
			(0.92)	
turkic:22			1.43	
			(1.38)	

	m2_full	m2_base	m2_lan	m2_TLD
turkic:23			1.09	
			(1.46)	
turkic:24			0.46	
			(1.49)	
TLD_au:1				−0.03
				(0.56)
TLD_au:2				0.06
				(0.38)
TLD_au:3				0.08
				(0.32)
TLD_au:4				0.09
				(0.29)
TLD_au:5				0.10
				(0.27)
TLD_au:6				0.11
				(0.26)
TLD_au:7				0.13
				(0.25)
TLD_au:8				0.14
				(0.24)
TLD_au:9				0.14
				(0.24)
TLD_au:10				0.15
				(0.24)
TLD_au:11				0.16
				(0.25)
TLD_au:12				0.17
				(0.26)
TLD_au:13				0.18
				(0.27)
TLD_au:14				0.19
				(0.29)
TLD_au:15				0.22
				(0.32)
TLD_au:16				0.28
				(0.36)
TLD_au:17				0.29
				(0.43)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_au:18				0.28
				(0.52)
TLD_au:19				0.18
				(0.73)
TLD_be:1				0.03
				(0.56)
TLD_be:2				0.03
				(0.40)
TLD_be:3				0.03
				(0.33)
TLD_be:4				0.03
				(0.30)
TLD_be:5				0.03
				(0.28)
TLD_be:6				0.05
				(0.27)
TLD_be:7				0.06
				(0.26)
TLD_be:8				0.08
				(0.25)
TLD_be:9				0.11
				(0.25)
TLD_be:10				0.12
				(0.25)
TLD_be:11				0.13
				(0.26)
TLD_be:12				0.14
				(0.27)
TLD_be:13				0.17
				(0.28)
TLD_be:14				0.19
				(0.30)
TLD_be:15				0.21
				(0.33)
TLD_be:16				0.29
				(0.39)
TLD_be:17				0.32
				(0.46)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_be:18				0.34
				(0.57)
TLD_be:19				0.45
				(0.85)
TLD_bg:1				−0.14
				(0.64)
TLD_bg:2				0.00
				(0.43)
TLD_bg:3				0.05
				(0.36)
TLD_bg:4				0.07
				(0.33)
TLD_bg:5				0.09
				(0.30)
TLD_bg:6				0.12
				(0.29)
TLD_bg:7				0.14
				(0.28)
TLD_bg:8				0.16
				(0.27)
TLD_bg:9				0.19
				(0.27)
TLD_bg:10				0.22
				(0.28)
TLD_bg:11				0.26
				(0.29)
TLD_bg:12				0.31
				(0.31)
TLD_bg:13				0.34
				(0.33)
TLD_bg:14				0.39
				(0.36)
TLD_bg:15				0.42
				(0.40)
TLD_bg:16				0.45
				(0.46)
TLD_bg:17				0.66
				(0.60)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_bg:18				0.53
				(0.71)
TLD_bg:19				0.74
				(1.12)
TLD_cy:1				−0.71
				(1.08)
TLD_cy:2				−0.03
				(0.59)
TLD_cy:3				0.06
				(0.49)
TLD_cy:4				0.02
				(0.45)
TLD_cy:5				0.12
				(0.41)
TLD_cy:6				0.09
				(0.39)
TLD_cy:7				0.20
				(0.38)
TLD_cy:8				0.32
				(0.37)
TLD_cy:9				0.49
				(0.38)
TLD_cy:10				0.70
				(0.41)
TLD_cy:11				0.80
				(0.45)
TLD_cy:12				1.05*
				(0.52)
TLD_cy:13				1.44*
				(0.66)
TLD_cy:14				1.20
				(0.66)
TLD_cy:15				1.40
				(0.82)
TLD_cy:16				22.97
				(35180.58)
TLD_cy:17				23.53
				(59036.12)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_cy:18				24.02
				(94775.28)
TLD_cy:19				24.23
				(157307.57)
TLD_cz:1				−0.07
				(0.59)
TLD_cz:2				−0.02
				(0.41)
TLD_cz:3				0.00
				(0.35)
TLD_cz:4				0.01
				(0.31)
TLD_cz:5				0.01
				(0.29)
TLD_cz:6				0.05
				(0.28)
TLD_cz:7				0.05
				(0.27)
TLD_cz:8				0.06
				(0.26)
TLD_cz:9				0.06
				(0.26)
TLD_cz:10				0.06
				(0.26)
TLD_cz:11				0.06
				(0.27)
TLD_cz:12				0.07
				(0.29)
TLD_cz:13				0.08
				(0.30)
TLD_cz:14				0.12
				(0.33)
TLD_cz:15				0.10
				(0.36)
TLD_cz:16				0.04
				(0.41)
TLD_cz:17				−0.03
				(0.49)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_cz:18				−0.01
				(0.60)
TLD_cz:19				0.14
				(0.93)
TLD_dk:1				−0.04
				(0.60)
TLD_dk:2				0.07
				(0.41)
TLD_dk:3				0.09
				(0.35)
TLD_dk:4				0.11
				(0.31)
TLD_dk:5				0.11
				(0.29)
TLD_dk:6				0.12
				(0.28)
TLD_dk:7				0.14
				(0.27)
TLD_dk:8				0.16
				(0.26)
TLD_dk:9				0.20
				(0.26)
TLD_dk:10				0.23
				(0.27)
TLD_dk:11				0.26
				(0.28)
TLD_dk:12				0.30
				(0.29)
TLD_dk:13				0.31
				(0.31)
TLD_dk:14				0.34
				(0.33)
TLD_dk:15				0.39
				(0.36)
TLD_dk:16				0.52
				(0.43)
TLD_dk:17				0.50
				(0.51)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_dk:18				0.63
				(0.65)
TLD_dk:19				0.55
				(0.92)
TLD_ee:1				−0.72
				(1.08)
TLD_ee:2				0.08
				(0.56)
TLD_ee:3				0.20
				(0.46)
TLD_ee:4				0.24
				(0.41)
TLD_ee:5				0.32
				(0.38)
TLD_ee:6				0.39
				(0.37)
TLD_ee:7				0.54
				(0.36)
TLD_ee:8				0.66
				(0.36)
TLD_ee:9				0.87*
				(0.38)
TLD_ee:10				1.04*
				(0.41)
TLD_ee:11				1.43**
				(0.49)
TLD_ee:12				1.63**
				(0.57)
TLD_ee:13				2.05**
				(0.75)
TLD_ee:14				2.52*
				(1.03)
TLD_ee:15				2.25*
				(1.03)
TLD_ee:16				1.93
				(1.04)
TLD_ee:17				1.54
				(1.05)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_ee:18				1.07
				(1.07)
TLD_ee:19				11.55
				(215.65)
TLD_es:1				−0.05
				(0.55)
TLD_es:2				−0.03
				(0.39)
TLD_es:3				−0.01
				(0.32)
TLD_es:4				−0.01
				(0.29)
TLD_es:5				0.00
				(0.27)
TLD_es:6				0.02
				(0.26)
TLD_es:7				0.04
				(0.25)
TLD_es:8				0.05
				(0.24)
TLD_es:9				0.07
				(0.24)
TLD_es:10				0.09
				(0.24)
TLD_es:11				0.11
				(0.25)
TLD_es:12				0.12
				(0.25)
TLD_es:13				0.14
				(0.27)
TLD_es:14				0.16
				(0.29)
TLD_es:15				0.19
				(0.31)
TLD_es:16				0.29
				(0.36)
TLD_es:17				0.31
				(0.42)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_es:18				0.36
				(0.52)
TLD_es:19				0.28
				(0.74)
TLD_fi:1				−0.10
				(0.61)
TLD_fi:2				0.04
				(0.41)
TLD_fi:3				0.09
				(0.34)
TLD_fi:4				0.10
				(0.31)
TLD_fi:5				0.10
				(0.29)
TLD_fi:6				0.12
				(0.27)
TLD_fi:7				0.14
				(0.26)
TLD_fi:8				0.17
				(0.26)
TLD_fi:9				0.20
				(0.26)
TLD_fi:10				0.22
				(0.26)
TLD_fi:11				0.25
				(0.27)
TLD_fi:12				0.26
				(0.28)
TLD_fi:13				0.27
				(0.30)
TLD_fi:14				0.31
				(0.32)
TLD_fi:15				0.34
				(0.35)
TLD_fi:16				0.45
				(0.41)
TLD_fi:17				0.62
				(0.51)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_fi:18				0.96
				(0.71)
TLD_fi:19				11.66
				(148.83)
TLD_fr:1				0.02
				(0.55)
TLD_fr:2				0.05
				(0.38)
TLD_fr:3				0.05
				(0.32)
TLD_fr:4				0.05
				(0.29)
TLD_fr:5				0.05
				(0.27)
TLD_fr:6				0.06
				(0.26)
TLD_fr:7				0.08
				(0.25)
TLD_fr:8				0.10
				(0.24)
TLD_fr:9				0.13
				(0.24)
TLD_fr:10				0.14
				(0.24)
TLD_fr:11				0.16
				(0.25)
TLD_fr:12				0.18
				(0.26)
TLD_fr:13				0.20
				(0.27)
TLD_fr:14				0.23
				(0.29)
TLD_fr:15				0.26
				(0.32)
TLD_fr:16				0.37
				(0.36)
TLD_fr:17				0.42
				(0.43)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_fr:18				0.49
				(0.54)
TLD_fr:19				0.77
				(0.85)
TLD_gr:1				−0.55
				(0.87)
TLD_gr:2				−0.06
				(0.52)
TLD_gr:3				−0.01
				(0.44)
TLD_gr:4				−0.03
				(0.40)
TLD_gr:5				0.01
				(0.37)
TLD_gr:6				−0.02
				(0.35)
TLD_gr:7				0.01
				(0.34)
TLD_gr:8				0.06
				(0.33)
TLD_gr:9				0.14
				(0.33)
TLD_gr:10				0.20
				(0.34)
TLD_gr:11				0.30
				(0.36)
TLD_gr:12				0.35
				(0.39)
TLD_gr:13				0.46
				(0.43)
TLD_gr:14				0.50
				(0.47)
TLD_gr:15				0.53
				(0.53)
TLD_gr:16				0.65
				(0.67)
TLD_gr:17				0.13
				(0.69)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_gr:18				0.71
				(1.07)
TLD_gr:19				−0.09
				(1.13)
TLD_hr:1				0.02
				(0.61)
TLD_hr:2				0.09
				(0.42)
TLD_hr:3				0.11
				(0.35)
TLD_hr:4				0.13
				(0.32)
TLD_hr:5				0.14
				(0.30)
TLD_hr:6				0.16
				(0.28)
TLD_hr:7				0.17
				(0.27)
TLD_hr:8				0.20
				(0.27)
TLD_hr:9				0.23
				(0.27)
TLD_hr:10				0.26
				(0.27)
TLD_hr:11				0.30
				(0.28)
TLD_hr:12				0.33
				(0.30)
TLD_hr:13				0.35
				(0.31)
TLD_hr:14				0.36
				(0.34)
TLD_hr:15				0.38
				(0.37)
TLD_hr:16				0.46
				(0.43)
TLD_hr:17				0.51
				(0.52)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_hr:18				0.55
				(0.65)
TLD_hr:19				0.46
				(0.92)
TLD_hu:1				−0.06
				(0.61)
TLD_hu:2				0.02
				(0.42)
TLD_hu:3				0.07
				(0.35)
TLD_hu:4				0.07
				(0.32)
TLD_hu:5				0.09
				(0.30)
TLD_hu:6				0.14
				(0.28)
TLD_hu:7				0.19
				(0.27)
TLD_hu:8				0.26
				(0.27)
TLD_hu:9				0.33
				(0.27)
TLD_hu:10				0.38
				(0.27)
TLD_hu:11				0.45
				(0.29)
TLD_hu:12				0.52
				(0.31)
TLD_hu:13				0.60
				(0.33)
TLD_hu:14				0.64
				(0.36)
TLD_hu:15				0.68
				(0.40)
TLD_hu:16				0.98
				(0.51)
TLD_hu:17				1.22
				(0.68)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_hu:18				1.77
				(1.06)
TLD_hu:19				12.86
				(295.53)
TLD_ch:1				−0.12
				(0.60)
TLD_ch:2				0.00
				(0.41)
TLD_ch:3				0.05
				(0.34)
TLD_ch:4				0.07
				(0.31)
TLD_ch:5				0.08
				(0.29)
TLD_ch:6				0.10
				(0.27)
TLD_ch:7				0.12
				(0.26)
TLD_ch:8				0.15
				(0.26)
TLD_ch:9				0.18
				(0.26)
TLD_ch:10				0.20
				(0.26)
TLD_ch:11				0.22
				(0.27)
TLD_ch:12				0.25
				(0.28)
TLD_ch:13				0.27
				(0.30)
TLD_ch:14				0.31
				(0.32)
TLD_ch:15				0.37
				(0.36)
TLD_ch:16				0.53
				(0.43)
TLD_ch:17				0.63
				(0.53)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_ch:18				0.59
				(0.65)
TLD_ch:19				0.50
				(0.92)
TLD_ie:1				−0.06
				(0.59)
TLD_ie:2				0.04
				(0.41)
TLD_ie:3				0.07
				(0.34)
TLD_ie:4				0.06
				(0.31)
TLD_ie:5				0.06
				(0.29)
TLD_ie:6				0.09
				(0.27)
TLD_ie:7				0.10
				(0.26)
TLD_ie:8				0.12
				(0.26)
TLD_ie:9				0.13
				(0.26)
TLD_ie:10				0.15
				(0.26)
TLD_ie:11				0.17
				(0.27)
TLD_ie:12				0.20
				(0.28)
TLD_ie:13				0.24
				(0.30)
TLD_ie:14				0.29
				(0.32)
TLD_ie:15				0.33
				(0.36)
TLD_ie:16				0.37
				(0.41)
TLD_ie:17				0.38
				(0.49)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_ie:18				0.34
				(0.60)
TLD_ie:19				0.06
				(0.79)
TLD_it:1				−0.12
				(0.55)
TLD_it:2				−0.08
				(0.38)
TLD_it:3				−0.05
				(0.32)
TLD_it:4				−0.04
				(0.29)
TLD_it:5				−0.03
				(0.27)
TLD_it:6				−0.01
				(0.25)
TLD_it:7				0.01
				(0.24)
TLD_it:8				0.02
				(0.24)
TLD_it:9				0.03
				(0.24)
TLD_it:10				0.03
				(0.24)
TLD_it:11				0.03
				(0.24)
TLD_it:12				0.04
				(0.25)
TLD_it:13				0.06
				(0.26)
TLD_it:14				0.08
				(0.28)
TLD_it:15				0.08
				(0.31)
TLD_it:16				0.11
				(0.35)
TLD_it:17				0.06
				(0.40)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_it:18				0.09
				(0.50)
TLD_it:19				0.06
				(0.71)
TLD_lt:1				0.07
				(0.69)
TLD_lt:2				0.17
				(0.47)
TLD_lt:3				0.16
				(0.40)
TLD_lt:4				0.19
				(0.36)
TLD_lt:5				0.19
				(0.34)
TLD_lt:6				0.24
				(0.32)
TLD_lt:7				0.26
				(0.31)
TLD_lt:8				0.33
				(0.31)
TLD_lt:9				0.45
				(0.31)
TLD_lt:10				0.56
				(0.32)
TLD_lt:11				0.67
				(0.34)
TLD_lt:12				0.75*
				(0.37)
TLD_lt:13				0.79*
				(0.40)
TLD_lt:14				0.91*
				(0.45)
TLD_lt:15				1.10*
				(0.53)
TLD_lt:16				1.59*
				(0.74)
TLD_lt:17				24.19
				(52107.73)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_lt:18				24.65
				(82845.37)
TLD_lt:19				24.88
				(138164.10)
TLD_lu:1				0.13
				(0.72)
TLD_lu:2				0.18
				(0.50)
TLD_lu:3				0.24
				(0.42)
TLD_lu:4				0.24
				(0.38)
TLD_lu:5				0.28
				(0.36)
TLD_lu:6				0.29
				(0.34)
TLD_lu:7				0.33
				(0.33)
TLD_lu:8				0.42
				(0.33)
TLD_lu:9				0.58
				(0.34)
TLD_lu:10				0.70*
				(0.35)
TLD_lu:11				0.87*
				(0.39)
TLD_lu:12				1.05*
				(0.43)
TLD_lu:13				1.11*
				(0.48)
TLD_lu:14				1.07*
				(0.52)
TLD_lu:15				0.79
				(0.53)
TLD_lu:16				1.40
				(0.76)
TLD_lu:17				1.00
				(0.77)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_lu:18				1.25
				(1.07)
TLD_lu:19				0.45
				(1.12)
TLD_lv:1				−0.35
				(0.80)
TLD_lv:2				−0.03
				(0.50)
TLD_lv:3				0.01
				(0.42)
TLD_lv:4				−0.02
				(0.38)
TLD_lv:5				−0.09
				(0.36)
TLD_lv:6				−0.07
				(0.34)
TLD_lv:7				−0.03
				(0.33)
TLD_lv:8				0.04
				(0.32)
TLD_lv:9				0.12
				(0.31)
TLD_lv:10				0.18
				(0.32)
TLD_lv:11				0.24
				(0.33)
TLD_lv:12				0.35
				(0.36)
TLD_lv:13				0.40
				(0.38)
TLD_lv:14				0.48
				(0.42)
TLD_lv:15				0.58
				(0.48)
TLD_lv:16				0.93
				(0.64)
TLD_lv:17				0.48
				(0.66)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_lv:18				0.45
				(0.80)
TLD_lv:19				0.36
				(1.12)
TLD_mt:1				−1.14
				(1.08)
TLD_mt:2				−0.29
				(0.55)
TLD_mt:3				−0.09
				(0.44)
TLD_mt:4				−0.04
				(0.39)
TLD_mt:5				−0.01
				(0.36)
TLD_mt:6				0.12
				(0.34)
TLD_mt:7				0.18
				(0.33)
TLD_mt:8				0.23
				(0.33)
TLD_mt:9				0.31
				(0.33)
TLD_mt:10				0.39
				(0.35)
TLD_mt:11				0.49
				(0.37)
TLD_mt:12				0.49
				(0.40)
TLD_mt:13				0.53
				(0.43)
TLD_mt:14				0.58
				(0.48)
TLD_mt:15				0.74
				(0.57)
TLD_mt:16				1.65
				(1.04)
TLD_mt:17				1.11
				(1.05)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_mt:18				23.91
				(85687.28)
TLD_mt:19				24.11
				(141517.51)
TLD_nl:1				0.02
				(0.57)
TLD_nl:2				0.08
				(0.40)
TLD_nl:3				0.09
				(0.33)
TLD_nl:4				0.09
				(0.30)
TLD_nl:5				0.09
				(0.28)
TLD_nl:6				0.11
				(0.27)
TLD_nl:7				0.13
				(0.26)
TLD_nl:8				0.15
				(0.25)
TLD_nl:9				0.18
				(0.25)
TLD_nl:10				0.21
				(0.25)
TLD_nl:11				0.24
				(0.26)
TLD_nl:12				0.27
				(0.27)
TLD_nl:13				0.31
				(0.29)
TLD_nl:14				0.34
				(0.31)
TLD_nl:15				0.42
				(0.35)
TLD_nl:16				0.58
				(0.41)
TLD_nl:17				0.76
				(0.51)

	m2__full	m2__base	m2__lan	m2__TLD
TLD__nl:18				0.85
				(0.65)
TLD__nl:19				1.29
				(1.12)
TLD__no:1				−0.02
				(0.60)
TLD__no:2				0.11
				(0.41)
TLD__no:3				0.12
				(0.35)
TLD__no:4				0.12
				(0.31)
TLD__no:5				0.12
				(0.29)
TLD__no:6				0.14
				(0.28)
TLD__no:7				0.17
				(0.27)
TLD__no:8				0.20
				(0.26)
TLD__no:9				0.23
				(0.26)
TLD__no:10				0.26
				(0.27)
TLD__no:11				0.30
				(0.28)
TLD__no:12				0.32
				(0.29)
TLD__no:13				0.34
				(0.31)
TLD__no:14				0.38
				(0.33)
TLD__no:15				0.43
				(0.37)
TLD__no:16				0.58
				(0.43)
TLD__no:17				0.74
				(0.53)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_no:18				0.69
				(0.65)
TLD_no:19				0.61
				(0.92)
TLD_other:1				0.04
				(0.39)
TLD_other:2				0.06
				(0.27)
TLD_other:3				0.07
				(0.23)
TLD_other:4				0.07
				(0.21)
TLD_other:5				0.09
				(0.19)
TLD_other:6				0.12
				(0.18)
TLD_other:7				0.13
				(0.17)
TLD_other:8				0.14
				(0.17)
TLD_other:9				0.15
				(0.17)
TLD_other:10				0.16
				(0.17)
TLD_other:11				0.18
				(0.17)
TLD_other:12				0.20
				(0.18)
TLD_other:13				0.22
				(0.19)
TLD_other:14				0.24
				(0.20)
TLD_other:15				0.26
				(0.22)
TLD_other:16				0.32
				(0.25)
TLD_other:17				0.35
				(0.29)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_other:18				0.38
				(0.35)
TLD_other:19				0.39
				(0.51)
TLD_pl:1				−0.04
				(0.57)
TLD_pl:2				−0.01
				(0.40)
TLD_pl:3				0.00
				(0.33)
TLD_pl:4				0.01
				(0.30)
TLD_pl:5				0.02
				(0.28)
TLD_pl:6				0.04
				(0.26)
TLD_pl:7				0.06
				(0.25)
TLD_pl:8				0.10
				(0.25)
TLD_pl:9				0.13
				(0.25)
TLD_pl:10				0.16
				(0.25)
TLD_pl:11				0.20
				(0.26)
TLD_pl:12				0.23
				(0.27)
TLD_pl:13				0.26
				(0.29)
TLD_pl:14				0.30
				(0.31)
TLD_pl:15				0.34
				(0.34)
TLD_pl:16				0.44
				(0.40)
TLD_pl:17				0.55
				(0.49)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_pl:18				0.79
				(0.65)
TLD_pl:19				1.24
				(1.12)
TLD_pt:1				−0.21
				(0.60)
TLD_pt:2				−0.10
				(0.41)
TLD_pt:3				−0.06
				(0.34)
TLD_pt:4				−0.04
				(0.31)
TLD_pt:5				−0.03
				(0.29)
TLD_pt:6				−0.01
				(0.27)
TLD_pt:7				0.01
				(0.26)
TLD_pt:8				0.02
				(0.25)
TLD_pt:9				0.04
				(0.25)
TLD_pt:10				0.05
				(0.25)
TLD_pt:11				0.07
				(0.26)
TLD_pt:12				0.08
				(0.27)
TLD_pt:13				0.10
				(0.29)
TLD_pt:14				0.11
				(0.31)
TLD_pt:15				0.11
				(0.33)
TLD_pt:16				0.14
				(0.38)
TLD_pt:17				0.16
				(0.46)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_pt:18				0.10
				(0.55)
TLD_pt:19				0.03
				(0.78)
TLD_ro:1				−0.12
				(0.56)
TLD_ro:2				−0.09
				(0.39)
TLD_ro:3				−0.07
				(0.33)
TLD_ro:4				−0.05
				(0.30)
TLD_ro:5				−0.05
				(0.28)
TLD_ro:6				−0.02
				(0.26)
TLD_ro:7				−0.03
				(0.25)
TLD_ro:8				−0.04
				(0.25)
TLD_ro:9				−0.08
				(0.24)
TLD_ro:10				−0.11
				(0.25)
TLD_ro:11				−0.14
				(0.25)
TLD_ro:12				−0.14
				(0.26)
TLD_ro:13				−0.13
				(0.28)
TLD_ro:14				−0.12
				(0.29)
TLD_ro:15				−0.15
				(0.32)
TLD_ro:16				−0.22
				(0.36)
TLD_ro:17				−0.26
				(0.43)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_ro:18				−0.23
				(0.53)
TLD_ro:19				−0.25
				(0.77)
TLD_se:1				−0.09
				(0.58)
TLD_se:2				0.00
				(0.40)
TLD_se:3				0.04
				(0.34)
TLD_se:4				0.04
				(0.30)
TLD_se:5				0.05
				(0.28)
TLD_se:6				0.06
				(0.27)
TLD_se:7				0.08
				(0.26)
TLD_se:8				0.09
				(0.25)
TLD_se:9				0.10
				(0.25)
TLD_se:10				0.11
				(0.25)
TLD_se:11				0.12
				(0.26)
TLD_se:12				0.13
				(0.27)
TLD_se:13				0.14
				(0.28)
TLD_se:14				0.15
				(0.30)
TLD_se:15				0.15
				(0.33)
TLD_se:16				0.19
				(0.38)
TLD_se:17				0.19
				(0.45)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_se:18				0.21
				(0.55)
TLD_se:19				0.20
				(0.79)
TLD_si:1				0.01
				(0.62)
TLD_si:2				0.06
				(0.43)
TLD_si:3				0.10
				(0.36)
TLD_si:4				0.13
				(0.33)
TLD_si:5				0.16
				(0.30)
TLD_si:6				0.17
				(0.29)
TLD_si:7				0.19
				(0.28)
TLD_si:8				0.22
				(0.28)
TLD_si:9				0.25
				(0.28)
TLD_si:10				0.25
				(0.28)
TLD_si:11				0.26
				(0.29)
TLD_si:12				0.23
				(0.31)
TLD_si:13				0.23
				(0.32)
TLD_si:14				0.28
				(0.35)
TLD_si:15				0.26
				(0.38)
TLD_si:16				0.33
				(0.45)
TLD_si:17				0.34
				(0.55)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_si:18				0.48
				(0.71)
TLD_si:19				0.69
				(1.12)
TLD_sk:1				0.07
				(0.60)
TLD_sk:2				0.07
				(0.42)
TLD_sk:3				0.15
				(0.35)
TLD_sk:4				0.13
				(0.32)
TLD_sk:5				0.12
				(0.30)
TLD_sk:6				0.15
				(0.28)
TLD_sk:7				0.19
				(0.27)
TLD_sk:8				0.22
				(0.27)
TLD_sk:9				0.25
				(0.27)
TLD_sk:10				0.28
				(0.28)
TLD_sk:11				0.31
				(0.29)
TLD_sk:12				0.36
				(0.30)
TLD_sk:13				0.43
				(0.33)
TLD_sk:14				0.52
				(0.36)
TLD_sk:15				0.64
				(0.41)
TLD_sk:16				0.89
				(0.53)
TLD_sk:17				1.52
				(0.83)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_sk:18				1.06
				(0.85)
TLD_sk:19				0.26
				(0.92)
TLD_uk:1				0.04
				(0.53)
TLD_uk:2				0.07
				(0.37)
TLD_uk:3				0.08
				(0.31)
TLD_uk:4				0.07
				(0.28)
TLD_uk:5				0.07
				(0.26)
TLD_uk:6				0.08
				(0.25)
TLD_uk:7				0.09
				(0.24)
TLD_uk:8				0.10
				(0.23)
TLD_uk:9				0.11
				(0.23)
TLD_uk:10				0.12
				(0.23)
TLD_uk:11				0.13
				(0.24)
TLD_uk:12				0.15
				(0.25)
TLD_uk:13				0.16
				(0.26)
TLD_uk:14				0.18
				(0.28)
TLD_uk:15				0.20
				(0.30)
TLD_uk:16				0.27
				(0.35)
TLD_uk:17				0.30
				(0.41)

	m2_full	m2_base	m2_lan	m2_TLD
TLD_uk:18				0.32
				(0.50)
TLD_uk:19				0.29
				(0.71)
Log Likelihood	−52246.13	−32606.77	−31067.25	−28161.34
DF	132733	74878	58896	41591
Num. obs.	132849	74936	59160	42199

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A.11: Statistical models

A.6 The TLD significance for Password-Password similarity

Table A.12: The statistical significance of TLDs for the Password-Password similarity

language	au	be	bg	cy	cz	dk	ee	es	fi	fr	gr	hr	hu	ch	ie	it	lt	lu	lv	mt	nl	no	other	pl	pt	ro	se	si	sk	uk
2																														
3																														
4																														
5																														
6																														
7																														
8																														
9							.																							
10							*											.												
11				.			*										.	*												
12				.			**										.	*												
13				*			**						.				*	*												
14				*			**						.				*	*												
15				.			*						.				*	*												
16				.			*						.				*													
17							.						.				*	.											.	
18													.																.	
19													.																	
20																														

A.7 BMA coefficients

Table A.13: BMA coefficients of the Model family 1

Variable	PIP	Post Mean	Post SD	Cond.Pos.Sign	Idx
Cyber	1	0.15	0.02	1	2
Literacy	1	1.48	0.05	1	3
Internet	1	-0.65	0.02	0	4
Democracy	1	0.17	0	1	5
PassLen	1	0.32	0	1	6
Effort2	1	-0.43	0	0	7
Effort3	1	0.09	0.01	1	8
Effort4	1	0.46	0.02	1	9
Mobile	0.01	0	0	1	1
SexF	0	0	0	1	10

Table A.14: BMA coefficients of the Model family 1 with sentiment

Variable	PIP	Post Mean	Post SD	Cond.Pos.Sign	Idx
Mobile	1	0.29	0.02	1	1
Cyber	1	1.05	0.03	1	2
Literacy	1	0.69	0.05	1	3
Internet	1	-0.91	0.03	0	4
Democracy	1	0.14	0	1	5
PassLen	1	0.32	0	1	6
Effort2	1	-0.41	0	0	7
Effort3	1	0.1	0.01	1	8
Effort4	1	0.5	0.02	1	9
SentPos	1	-0.34	0.05	0	11
SentNeg	1	0.85	0.05	1	12
SexF	0.03	0	0	0	10

Table A.15: BMA coefficients of the Model family 2

Variable	PIP	Post Mean	Post SD	Cond.Pos.Sign	Idx
Cyber	1	-0.37	0.03	0	1
Mobile	1	-0.22	0.02	0	2
Literacy	1	1.16	0.09	1	3
Internet	1	-1.35	0.05	0	4
Democracy	1	0.02	0	1	5
SexF	1	0.1	0.01	1	6