UNIVERZITA KARLOVA

FAKULTA SOCIÁLNÍCH VĚD

Institut sociologických studií

Katedra sociologie

Bakalářská práce

2021 **Viktor Jurdič**

UNIVERZITA KARLOVA

FAKULTA SOCIÁLNÍCH VĚD

Institut sociologických studií

Katedra sociologie

Tematické modelování v analýze facebookových příspěvků: Politici a pandemie koronaviru

Bakalářská práce

Autor práce: Viktor Jurdič

Studijní program: Sociologie a sociální antropologie

Vedoucí práce: doc. Mgr. Martin Hájek, Ph.D

Rok obhajoby: 2021

Prohlášení

1. Prohlašuji, že jsem předkládanou práci zpracoval samostatně a použila jen uvedené prameny a literaturu.
2. Prohlašuji, že práce nebyla využita k získání jiného titulu.
3. Souhlasím s tím, aby práce byla zpřístupněna pro studijní a výzkumné účely.

V Praze dne 04. 05. 2021 Viktor Jurdič

Bibliografický záznam

JURDIČ, Viktor. *Tematické modelování v analýze facebookových příspěvků: Politici a pandemie koronaviru.* Praha, 2021. 40 s. Bakalářská práce (Bc). Univerzita Karlova, Fakulta sociálních věd, Institut Sociologických studií. Katedra Sociologie. Vedoucí bakalářské práce doc. Mgr. Martin Hájek Ph.d

Rozsah práce: 57 994 znaků (od úvodu po závěr)

Anotace

Tato bakalářská práce diskutuje využití tematického modelování, konkrétně strukturálního tematického modelu při využití tematické analýzy na datech sestávajících z facebookových příspěvků politiků v rozsahu jednoho roku. Dále se tato práce zabývá využitím tematického modelování jako první fáze výzkumu s následnou kvalitativní analýzou textů. V textech příspěvků jednotlivých politiků z roku 2020 poznamenaného pandemií koronaviru bylo tímto postupem odhaleno, jak tento fenomén tematicky rámují. Tato práce nepřináší definitivní odpověď na otázku, zda je pomocí tematického modelování vhodné provádět výběr textů pro kvalitativní analýzu v mixed method paradigmatu. Výsledek analýzy ale ukazuje, že tento postup může při použití metodologického rámce na vhodných datech podávat validní výsledky.

Annotation

This bachelor thesis discusses an application of structural topic model on the data of facebook posts of politicians from a year long period of the year 2020 which was dramaticly affected by the pandemic of coronavirus. Further more this thesis proposes a mixed method framework consisted of follow up qualitative analysis of for the topic important posts. This followup was succesfully performed on a sample of data, providing an insight view over topic framing of covid-19 during the pandemic. Based on the aplication of method, this thesis do not give a clear answer whether topic modeling is a fitting tool for topic selection of significant posts. Resuts of the analysis show the possibility application of the framework when used on suitable data.

Klíčová slova

tematické modelování, mixed method, sociální sítě, politická komunikace, covid-19

Keywords

topic modeling, mixed method, social media, political communication, covid-19

Title

Topic modeling of Facebook posts: Politicians and Coronavirus pandemic

Poděkování

Chci poděkovat vedoucímu své bakalářské práce doc. Martinu Hájkovi za podnětné rady a nasměrování, které mě přivedlo k tematickému modelování jako tématu mé práce.

1 Obsah

1. Texty, sociální sítě a společenské vědy

Sociolog, politolog nebo snad antropolog má dnes možnost sáhnout do rozsáhlé knihovny metod analýzy textu a zvolit takovou, která nejlépe přilne k výzkumné otázce i analyzovaným datům. Výzkumník, cestovatel v čase, který by navštívil knihovnu před vynálezem knihtisku, by si zřejmě vystačil se základními kvalitativními metodami a omezené množství textů by pro něj bylo uchopitelným materiálem. Dnes si ale můžeme pokládat mnoho výzkumných otázek, jejichž zodpovězení by bylo skutečnou výzvou za použití klasických nástrojů analýzy textu, jejichž výčet nalezneme v každé učebnici základů sociologických metod. Limit dnes tvoří zejména podoba dat. S rozvojem internetu a na pozadí technologického a ideologického základu nazývaného Web 2.0, jenž je založený na obsahu, který vytváří sami uživatelé (Kaplan a Haenlein 2010), dnes vzniká nepřeberné množství textů, které reflektují a utvářejí sociální realitu. Zejména sociální sítě jsou pak prostředím, které je prakticky nevyčerpatelnou knihovnou textů.

Rozvoj informačních technologií, který přivedl stovky milionů lidí do prostředí sociálních sítí, dal rovněž sociálním vědcům do rukou nástroje, jak nové prostředí efektivně zkoumat. Digitální data se s počítačovou výpočetní silou potkávají na poli mnoha společenskovědních disciplín. Výzkum komplexních a rozsáhlých sociálních sítí ve smyslu vazeb ve společnosti, sociologie vědy v dosavadním rozsáhlém vědeckém poznání, sociální psychologie, studium reprodukce kultury i politická sociologie jsou jedny z oblastí sociologie , které počítačovou analýzou dat otevírají nové možnosti poznání (Edelmann et al. 2020).

1.1 Sociologická analýza textu

Sociologie a další společenskovědní disciplíny, které s touto vědou sdílí metody výzkumu, často využívají text jako materiál, který analyzují. Ať už se jedná o přepis rozhovoru nebo text, který vždy existoval jen na papíře, nebo jako jedničky a nuly v počítačové paměti, existuje mnoho různých způsobů, jak k analýze textu přistoupit.

Pohled na to, jak chápat tuto rozmanitost metod nabízí mnoho autorů. Bernard a Ryan (1997) spatřují dvě základní tradice: lingvistickou a sociologickou. Pro lingvistickou analýzu je text vlastním předmětem výzkumu a je spojena především s antropologickým bádáním. Sociologický přístup chápe text jako „okno do lidské zkušenosti“ (595). Autoři v textu

představují paletu metod analýzy textu. Grounded theory, různé formy formální obsahové analýzy založená na induktivním ale i deduktivním kódování i strukturní analýzu a analýzu sémantické sítě, které text transformují do kvantitativní podoby. Další obdobnou metodou jsou kognitivní mapy, které „kombinují intuici lidských kodérů s kvalitativní metodou síťové analýzy“ (621).

Obdobnou škálu metod analýzy textu představuje Kronick (1997). Popisuje tři základní druhy formální textové analýzy, které sociologie využívá k analýze textu. Formální obsahová analýza je kvantitativní metodou, která využívá principu „konverze verbálního textu na měřitelné proměnné“ (58) a má za úkol testovat definované hypotézy. Oproti tomu metoda zakotvené teorie je nástrojem, který byl vytvořen za účelem tvorby teorie. Hermeneutická analýza vyžadující dobrou znalost kontextu má pak kořeny v humanitárních vědách a společenské vědy ji jako nástroj spíše přebírají.

Dvojice autorů Bolden a Moscarola (2000)se v článku prozaicky nazvaném *Bridging the Quantitative-Qualitative Divide* zabývají využitím lexikální analýzy jako nástroje, který je možné využít při analýze velkého objemu textových dat. Tato analýza umožňuje nejen odhalit sémantický obsah, ale dovede odhalit také strukturní rysy jazyka. Autoři zároveň uvádějí, že odlišnost výpovědí je možné hledat v rozdílném použití slovní zásoby, a to i v případech, kdy je obsah výpovědi stejný. Jako příklad zde slouží rozdílná odbornost mluvčích, s kterou se mění i lexikální prostředky.

Mnohé limity metod řeší výzkumníci takzvaným mixed method designem, ve kterém může nedostatky metody jedné vyvážit doplňující odlišná metoda. Snelson (2016) představuje metodologickou metaanalýzu studií z období 2007-2013, které se věnovaly výzkumu sociálních médií a u kterých bylo použito kvalitativního nebo mixed method přístupu. Celkově bylo analyzováno 229 studií, z toho 55 kombinovalo kvalitativní a kvantitativní techniky. Tento podíl ukazuje, že paradigma mixed method je pro výzkum sociálních sítí užívaný nástrojem.

Macek (2012) se v článku *Nevyčíslitelné porozumění: Kvalitativní výzkum online sociálních sítí* věnuje limitům, které může mít poznání odkázané pouze na kvantitativní metody analýzy textu. Autor argumentuje, že s rozvojem sociálních sítí dochází ke změnám v sociálním chování v kyberprostoru. Induktivní metody jsou tak vhodné jednak protože neexistuje

dostatečné množství teorií, ale také protože mohou zachycovat směr samotných změn. Autor v článku definuje kvantitativní výzkum jako výzkum, který „pracuje s číselnými vlastnostmi sociální reality a s numerickým kódováním jinak nenumerických proměnných. Jedná se tedy o takový výzkum, který pozorované skutečnosti převádí do podoby, jež je dále zpracovatelná matematickými/statistickými metodami“ (139). Kvalitativní výzkum je popsán jako výzkum který se ptá na to „jak a proč sociální aktéři činí a jak svému jednání rozumí a jak je reflektují (neboli jaké významy v souvislosti se svým jednáním produkují a naopak jaké významy jejich jednání strukturují), a nebo tím, jaké významy nesou zkoumané texty a jak jsou tyto významy textem utvářeny“(140). Macek zmiňuje, že časté výhrady vůči kvalitativnímu výzkumu jsou jeho „měkkost“, kdy je důležitý význam, ne vztahy proměnných, a malý počet případů, a z toho plynoucí obtížné zobecnění.

Premisou tak zůstává, že strojové čtení textu analyzuje pouze textový povrch. Lidský čtenář pak do čtení promítá svou pozici sociálního aktéra a do jeho analýzy také vstupuje předporozumění tématu (Hájek 2014).

Aparát metod, které sociologický výzkum využívá, reaguje na změny v povaze dat – dostupných textů. Dochází k rozvoji počítačové analýzy textu, která si umí poradit s velmi dlouhými texty i soubory textů, aniž by vyžadovala zásah lidského čtenáře během první fáze analýzy textu. V této sféře se významně uplatnily metody tematického modelování.

1.2 Tematické modelování

Sociologické nástroje zpracování textu pracují s jednou společnou myšlenkou, a tou je redukce komplexity analyzovaného text. Kódování lidským čtenářem si klade za cíl sloučit různá vyjádření pod jednotlivé kódy, do stejných skupin, a umožnit zobecnění samotné analýzu textu. Stejnou myšlenku následuje i zpracování textu pomocí počítače. Výpočetní technika je schopná rychle přečíst velké množství textu a zaznamenat výskyt konkrétního slova v textu. Umožňuje tak komplexní texty redukovat do datových matic. Souvislí text se pro mění do položky, u které je zaznamenán výskyt konkrétních slov. Takto zjednodušený text je pak možné podrobovat sofistikovaným statistickým metodám. Tuto myšlenku na počátku 80. let minulého století využila metoda latentní sémantické analýzy (LSA) schopná zařadit text do tematické skupiny, později navazuje pravděpodobnostní latentní sémantická analýza (pLSA) (Crain et al. 2012a).

Právě poslední zmíněnou metodou se otevírá soubor metod tematického modelování. Analýza pomocí tematického modelování je induktivní metodou, která se v kontextu strojové analýzy textu řadí mezi nejkomplikovanější, ale také mezi metody pracující na principu bayesianské pravděpodobnosti, které mají potenciál stát se běžnou metodou pro automatickou analýzu textu (Stieglitz a Dang-Xuan 2013). Jedná se o soubor metod přirozeného zpracování jazyka, která našla bohaté uplatnění v analýze sociálních sítí nebo vyhledávání informací (Jelodar et al. 2019).

Tyto metody mají mnoho společného se shlukováním a redukcí dimenzí. Všechny tyto metody se zaměřují na strukturu vztahů mezi texty, přesto jsou odlišné. Shlukování přisuzuje dokumentu příslušnost do shluku na základě podobnosti nebo nepodobnosti s ostatními dokumenty. V případě měkkého shlukování (soft clustering) je dokument zařazen do více shluků. Uvažováním shluků jako dimenzí vzniká nízko-dimensionální reprezentace dokumentů. Redukce dimenzí naopak nejprve identifikuje klíčové vlastnosti dokumentu, a následně snižuje počet dimenzí, tak aby výsledné dimenze vystihovali původní reprezentaci. Je tak zachováno větší množství původních informací, protože nové dimenze jsou často odvozeny od předchozích, je vyvození závěrů stále komplikované. Tematické modelování skládá oba předchozí principy. Dokumentu jsou připsána latentní témata založená na shluku dokumentu i vlastnosti v rámci celého datového souboru. Výsledkem je přesná identifikace vlastností dokumentu v rámci celého souboru a zachování použitelného množství informací pro vytvoření závěrů o jeho povaze. (Crain et al., 2012: 131-132)

> „V uplynulých letech došlo k vývoji a úspěšné aplikaci několika latentních faktorových modelů na diskrétních datech. Jedním z významných příkladů je Hofmannova pLSI (aspect model). Tato metoda si získala pozornost mnoha výzkumníků a byla aplikována v modelování textu, kolaborativním filtrování a link analysis. V kontextu modelování textu je pLSI typ modelu „balík slov (bag-of-words)“, který přehlíží pořádek slov v dokumentu. Metoda je založena na redukci dimenzí, která umístí každý dokument do místa v nízko-dimensionálním prostoru ‚témat‘.“(Blei et al. 2002)

Latentní Dirichletovu alokaci představuje trojice autorů Blei, Ng a Jordan v roce 2003 v *Journal of Machine Learning Research.* Již o rok dříve, ale v rámci konference *Advances*

in Neural Information Processing Systems 14 představili a publikovali, ne tak podrobný, avšak v klíčových částech prakticky totožný text.

> „V LDA předpokládáme, že existuje **k** skrytých latentních témat, podle kterých jsou vytvářeny texty dokumentů, a že každé téma je reprezentováno jako multinomiální distribuce **|V|** slov ve slovníku. Dokument je generován použitím některých těchto témat a následným použitím některých slov z daných témat“ (Blei et al. 2002: 602).

Stejně jako se metoda vyvinula zdokonalením dříve existujících méně komplexních metod, docházelo i po jejím vzniku k rozvoji. Nad možnostmi se zamýšlí mimo jiné autor metody David Blei. V článku *Probabilistic Topic Models* se mimo jiné zabývá možnými aspekty LDA, které je možné rozvíjet. Upozorňuje například na možnost nepřistupovat k dokumentu pouze jako k neuspořádanému balíku slov a zmiňuje tematický model, který bere v úvahu také pořadí slov v textu. Dále uvádí možnost zohlednit pořadí samotných dokumentů, v případech, kdy jsou data sbírána po dlouhou dobu let, a dá se tak očekávat vývoj tématu. Pro takové situace vznikl dynamický tematický model. Vznikla také neparametrická verze LDA, v které výzkumník nemusí předpokládat počet témat. Tato metoda byla následně obohacena o hierarchickou složku. Sférický tematický model zase umožňuje identifikovat také pro téma nepravděpodobná slova. Možnost zahrnout k textům metadata, používat LDA na jiných druzích dat, nebo možnosti vizualizace výsledků výpočtů jsou oblasti v který autor původní metody spatřuje prostor pro další rozvoj. (Blei 2012)

Tematické modelování na základě LDA se dále rozvíjí. Zejména vznikají modely, které kombinují LDA a další metody a cílí tak na snížení počtu dokumentů nutných k analýze, nebo analýzu krátkých textů (Nguyen et al. 2015), analýza je doplňována o další roviny, jako je hierarchizace témat (Wallach 2006), ale vznikají také algoritmy nové jako BTM, jehož cílem je lépe se vypořádat s krátkými texty (X. Cheng et al. 2014). I na tento model již vznikly adaptace (X. Wu a C. Li 2019), případně jeho kombinací s jinými metodami je cílem kombinovat jejich kvality (L. Li et al. 2018) (Nguyen et al. 2015) (Q. Liqing et al. 2019). To ukazuje rychlý a také pohotový vývoj, kdy je metoda upravována tak, aby vyhovovala změně v povaze zkoumaných dat, která s rozvojem technologií a sociálních sítích probíhá. Zejména je patrná snaha vytvořit metodu která by podávala dobré výsledky při analýze krátkých textů *Shotr Text Topic Modeling* (Qiang et al. 2018) svým názvem tuto snahu potvrzuje.

K tematickému modelování se používají programovací jazyky jako je R nebo Python. To je flexibilní a vhodnou formou zpracování i přes vyšší vstupní nároky na znalost tohoto nástroje (Nelson 2020). Na základě toho můžeme předpokládat, že i v důsledku ukotvení tematického modelování v R nebo Pythonu vznikají varianty metody uzpůsobené pro různá data i výzkumné cíle.

Dochází i k menším úpravám LDA. Příkladem jsou hierarchická LDA (hLDA) která identifikuje i podtémata a vytváří hierarchickou strukturu (Crain et al. 2012a), nebo topic-link LDA, která umožňuje sledovat vazbu mezi tématy (Liu et al. 2009). Vzájemné vazby témat umožňuje sledovat i strukturní tematický model (STM).

1.3 Strukturální tematický model

Strukturální tematický model vychází z LDA a je jeho obohacenou variantou, která umožňuje při analýze textů využít i další proměnné (Roberts et al. 2013). Roberts a další (2016) zdůrazňují, že sociální vědy povětšinou pracují s daty, která obsahují i další informace než pouze samotný text, který chceme analyzovat. Často navíc předpokládáme souvislost výskytu témat v textu a hodnot dalších proměnných, které k textu přináleží. V simulované analýze pak ukazují, že po zahrnutí další dostupné proměnné je model schopný lépe vystihnout příslušnost textu k tématům.

Vliv další proměnné na STM může být dvojího charakteru. Hodnota proměnné může ovlivňovat pravděpodobnost, s kterou model předpokládá výskyt tématu v textu, nebo může mít vliv na slovní zásobu, kterou je téma definováno (Lucas et al. 2015). Tamtéž je zdůrazňována další významná vlastnost STM, kterou je možnost odhalit korelaci témat v textech. Model je schopný ukázat společný výskyt témat v textech, a výzkumu se tak otevírá možnost popisovat vztahy mezi jednotlivými tématy

1.4 Aplikace tematického modelování

Tematické modelování našlo svou neoddiskutovatelnou roli v analýzách vědeckého poznání, kde umožňuje strojově přečíst abstrakty vědeckých článků a určit témata, kterým se věnují (Edelmann et al. 2020). Například Cheng et al. (2020) nabízejí použití LDA pro zmapování odborné literatury na téma nemoci covid-19 a srovnání MERS a SARS.

V oblasti sociálních sítí byl za použití LDA proveden výzkum komunikace na Twitteru prezidentských kandidátův USA pro volby v roce 2016 (Ryoo a Bendle 2017). Tento výzkum navíc využil dynamického modelu a sledoval, jak se témata kandidátů mění v čase. Z prostředí Facebooku vznikl studie, která analyzovala jaká témata na sociální síti komunikují muži a jaká ženy, a také jaká je míra reakcí na tyto příspěvky (Wang et al. 2013). V Německu proběhla studie, jejímž cílem bylo analyzovat rozdíl v politickém diskurzu v facebookových příspěvcích extremistických a ostatních politických stran (Stier et al. 2017). Do analýzy pomocí LDA vstoupilo 244 tis. příspěvků.

Tematický model může být použit v e spolupráci s kvalitativní metodou. Rodriguez a Storer (2020) využívají STM jako druhé fáze výzkumu obsahu Twitterových příspěvků. 3060 tweetů bylo nejprve induktivně kódováno, tento datový korpus byl pak tematicky modelován. V závěru bylo provedeno semantincké srovnání kódů, které ukazuje, že metody se vhodně doplňují. Hlavní výhodou tematického modelování je schopnost poskytnout přehled analyzovaných dat.

Právě této charakteristiky využívá postup počítačové zakotvené teorie.

1.5 Počítačová zakotvená teorie

Metodologický přístup využívající kombinaci tematického modelování a analýzy lidským čtenářem navrhuje Laura K. Nelson (2020) v článku *Computational Grounded Theory.* Postup analýzy založený na třech krocích nejprve využívá počítačovou analýzu textu. Jednou z možných metod v tomto kroku je tematické modelování, které vytvoří kategorie, tematické celky, a úlohou výzkumníka je kategorie pouze interpretovat. Tato vrstva přispívá k induktivnímu charakteru metody a k eliminaci vlivu předpokladů výzkumníka. V uvedené aplikaci je použit strukturální tematický model pro analýzu rozdílu v dokumentech produkovaných feministickým hnutím. Možností STM je využito pouze do té míry, že je modelování rozděleno podle několika známých metadat textů: místa a času vzniku, respektive města a vlny feministického hnutí. Tento krok umožňuje výzkumníkovi rychle se seznámit se strukturou a povahou dat.

V druhém kroku, který vyžaduje lidského čtenáře, se výzkumník vrací k textům a analyzuje je kvalitativně. Pro zmenšení objemu analyzovaného textu je vhodné využít struktury

odhalené v předchozím kroku. Cílem této fáze je upřesnit hypotézy formulované v předchozím kroku, případně vyvrátit některá milná zařazení textů k nevhodným tématům, kterých se počítačový algoritmus může dopustit, protože nezohledňuje jazykové prostředky, jako je humor, ironie či metafora. V případě aplikace na analýzu feministického hnutí, první krok definuje nejčastěji vyskytující se témata v textech a dokumenty s největším výskytem těchto témat. Kvalitativní analýze v druhém kroku je podrobeno deset dokumentů s nejvyšším výskytem každého z dvanácti nejčastějších témat.

Výše popsaný postup, který se po vzoru zakotvené teorie volně pohybuje mezi oběma kroky a po čtení textu výzkumníkem v druhém kroku dochází k revizi tematických struktur textů vytvořených v prvním kroku, uzavírá Nelson krokem třetím, který má ověřit dosavadní závěry. Tento krok vyžaduje operacionalizaci vzorců, identifikovaných v textech v předchozích krocích. V případě příkladového využití metody autorkou, jsou ověřovány vzorce o využívání obecnějších slov v jednom podsouboru textů a slov konkrétnějších v podsouboru druhém a častější zmínku o organizacích v jenom podsouboru oproti jedincům v druhém podsouboru. K tomuto úkonu je použita lexikální databáze WordNet. Taková databáze je dostupná i pro český jazyk (Pala et al. 2010) a v případě obdobných výzkumných cílů, by byl celý postup *počítačové zakotvené teorie* ve výše popsané podobě aplikovatelný na texty v českém jazyce.

Data o použitá ve výše zmíněné aplikaci metody čítají 1023 textů (Nelson 2017) a zjištění kvantitativně ověřovaná v posledním kroku jsou spíše obecná a do této části analýzy tak vstoupil datový korpus prakticky v celém rozsahu.

1.6 Koronavirus a politika na sociálních sítích

Lidé jsou koronavirem a reakcí společnosti na jím způsobované onemocnění fyzicky rozděleni. Kontakt jim umožňují sociální sítě a tohoto prostředí také chodí čerpat informace. Přestože jsou právě sociální sítě médiem, na kterém se rapidně šíří dezinformace, využívají jich ke komunikaci témat spojených s koronavirem i politické subjekty. Pro tento účel sociálních sítí hojně využívala administrativa bývalého prezidenta USA Trumpa. V tomto případě však k oběma zmíněným účelům. (Limaye et al. 2020)

Roli politiků a jejich vliv na informovanost obyvatel o novém fenoménu jakým byl koronavirus není možné podceňovat. Ve Spojených státech sehráli kromě Světové zdravotnické organizace klíčovou roli prezidenti Barack Obama a Donald Trump (Yum 2020). Pro utváření významu informací skrze sociální sítě, které v obdobích krize narůstá, jsou klíčoví aktéři, kteří předávají velké množství informací (Mirbabaie et al. 2020). Právě politici mohou zastávat roli informačních uzlů v prostředí sociálních sítí.

Podobu informací na sociálních sítích mohou politici ovlivňovat nejen rozhodnutím, zda se budou nebo nebudou podílet na šíření konkrétních informací, případně dezinformací. Mohou prostřednictvím šířených informací rámovat podobu témat. Ve Spojených státech byl koronavirus, respektive covid-19, tématem zdravotní péče a veřejného zdravý pro demokraty, ale tématem ekonomickým, spojeným s krizí drobného podnikání pro republikány (Panda et al. 2020). Tato rámování a interpretace pak přebírají konzumenti obsahu sociálních sítí politiků.

Ať už by politici téma koronaviru v prostředí sociálních sítí přijali za své či nikoliv, bude covid-19 i fenoménem politickým. Podoba dopadů onemocnění covid-19 na společnost je řízena politickým rozhodnutím. K politikům proto i z akademické obce míří apel nejen k prováděným krokům v těchto souvislostech, ale i ke společné komunikaci problému, někdy implicitní, jindy vyřčený (Murdoch et al. 2020).

Tento apel je na místě. Podoby politické komunikace a rámování témat ve veřejném prostoru nemáí dopad pouze na informace samotné, ale ovlivňují chování jednotlivců i celé společnosti (Barrios a Hochberg 2020).

2 Výzkumné cíle

„Identifikovat témata, která se v textu skrývají, je pro kvalitativní výzkum zásadní, nicméně je to pouze prvním krokem výzkumu. Interpretace témat je klíčovým bodem analýzy. Přímo aplikovat automatickou analýzu textu je velmi efektivní pro získání rozkrytí tematických struktur dat, ale referovat přímo k výsledkům modelu nechává výzkumníka v nebezpečí přílišného zjednodušení… Použití mixed method výzkumný design může pomoci zacelit slabiny každého z přístupů.“ (Chakrabarti a Frye 2017: 1357)

Počátkem roku 2020 se po celém světě rozšířila pandemie nemoci covid-19, která zasáhla nejen každodenní život, ale stala se také politickým tématem. Odpovědět na otázku, jak se téma covid-19 promítlo do komunikace politiků na Facebooku, je výzkumným cílem této práce. Mým cílem bylo odděleně analyzovat několik souborů dat textů facebookových příspěvků politiků za rok 2020.

K tomu jsem využil tematického modelování pomocí strukturálního tematického modelu. Mým cílem je obdobně jako u výše popsaného postupu počítačové zakotvené teorie využít tohoto prvního kroku k zúžení objemu dat na takové texty, které jsou pro pochopení rámování koronaviru nejpodstatnější.

V této práci chci odpovědět na otázku, jak čeští politici konstruovali téma covid-19 v roce 2020 a v souvislosti s jakými aspekty toto téma rámovali. Pomocí toho chci zjistit, jak se STM chová na obdobných datech, tj. rozsahem nepřesahujícím 800 textů v korpusu a analyzovaní politici jsou zvoleni tak, aby jednotlivé texty vykazovaly různé charakteristiky. Někde například s výskytem i velmi dlouhých textů, jinde se stejně dlouhými texty.

Tato práce má za cíl ověřit použitelnost STM jako nástroje pro analýzu textů na datech s menším rozsahem, ale i mixed method designu STM a kvalitativní analýzy, jako výzkumného postupu.

3 Metodologická část

3.1 Data

Cílem bylo vytvořit několik datových korpusů, obsahující texty facebookových příspěvků politiků, které by mohly vstoupit do analýzy politické komunikace na této sociální síti.

Jak ukazuje Macková (2018, 105), byl mezi lety 2013-2015 nejpopulárnějším komunikačním médiem českých poslanců Facebook. Přestože sociální média se rychle vyvíjí, je Facebook stále častým komunikačním prostředkem českých politiků. Proto volím prostředí této sociální sítě pro získání dat.

První fází sběru dat byl výběr vhodných profilů politiků na Facebooku. Tento výběr byl prováděn z oficiálních profilů politiků působících v celostátní politice tak, aby bylo možné

výsledky analýz jednotlivých politiků následně srovnat. Zásadním kritériem bylo množství příspěvků na daném profilu. Zejména pak příspěvků autorských, případně sdílených příspěvků, které byly doplněny vlastním textem. Tato orientace v potenciálních datech probíhala na počátku února 2021 s předpokladem, že se četnost přispívání na daném profilu zásadně neproměnila.

Pro následnou fázi, kterou byl samotný sběr dat, jsem využil programu Facepager (Jünger a Keyling 2019) ve verzi 4.3.9. Tento nástroj umožnuje stahovat data ze sociálních sítí jako Twitter, YouTube nebo právě Facebook. V případě Facebooku je zapotřebí v nástroji zadat identifikační číslo veřejného profilu, ze kterého chceme data stahovat. Ze všech profilů však nejsou data tímto nástrojem dosažitelná a nastavením facebookového účtu je možné použití Facepageru znemožnit. Z potenciálně vhodných profilů předsedy České pirátské strany Ivana Bartoše (Ivan Bartoš) a poslance stejné strany Ondřeje Profanta (Ondřej Kedrigern Profant), tak nebylo možné data pomocí nástroje Facepager získat.

V prostředí Facepageru je nutné zvolit strukturu požadovaných dat. *Message* vrátí autorský text příspěvku, *created_time* čas a datum, kdy byl příspěvek zveřejněn a *type* identifikuje případné další prvky příspěvku, jako je přiložený obrázek, video, odkaz, nebo sdílený příspěvek. Automaticky je součástí dat *object_id,* které je unikátním identifikátorem příspěvku. Výsledná data obsahují právě tyto čtyři proměnné. Další proměnné, které Facepager generuje automaticky budou odstraněny.

Následné zpracování dat probíhá pomocí programovacího jazyka R v programu R Studio, který nabízí uživatelsky přívětivější prostředí, než je příkazový řádek v případě využití R bez grafického rozhraní. Prostředí je tak podobné tradičnímu softwaru kvantitativní analýzy, jako je například SPSS.

Již v prvním kroku, kterým je import dat do prostředí R Studia, je na místě ostražitosti. Pozornost je třeba věnovat kódování znaků v získaných datech. V případě analýzy textů v českém jazyce může dojít například ke ztrátě diakritiky, v horším případě způsobí záměna znaků volbou jiného kódování úplnou nečitelnost textu. V případě použití programu Facepager jsou získaná data ve formátu UTF-8 a při importu je tak nutné zvolit právě toto kódování znaků.

Importovaná data byla transformována do podoby *data frame.* Dále byly vytvořeny proměnná s datem publikování příspěvku a spojitá proměnná den v roce.

3.1.1 Úprava dat pro tematické modelování

Příspěvky na sociálních sítích nejsou strohým textem, na který by bylo možné okamžitě aplikovat tematické modelování. Texty příspěvků v analyzovaných datech obsahují 980 unikátních znaků. 41 znaků české abecedy včetně písmen s diakritickými znaménky je v textech obsaženo i v podobě velkých písmen a v textech se vyskytují také samotná diakritická znaménka. Na tomto vysokém počtu znaků se podílí také písmena z cizích abeced, která jsou výsledkem odkazů politiků na zahraniční osobu nebo organizaci. Vzhledem ke globální epidemii nebyly výjimkou ani čínské znaky. Zdaleka největší množství znaků pak tvoří emojies. Tyto znaky, jako například (👍 😂 😠) jsou unikátními znaky, a liší se tak od emoticonů (:D ;) :O), které jsou složení z jednotlivých znaků interpunkce.

Texty je proto nutné upravit do podoby podoby, kdy budou lépe počítačově čitelné. Emojies a znaky z cizích abeced byly odstraněny v kroku lemmatizace. Lemmatizece je proces, během kterého jsou slova převedena na svůj základní tvar. Výzkumník pracující s daty v českém jazyce však v tomto kroku naráží na komplikaci. Programovací jazyk R nenabízí nástroj pro lemmatizaci v českém jazyce. Proto jsem k tomuto kroku využil nástroj MorphoDiTa[1] v jeho online podobě při použití jazykového modelu[2] založeného na morfologickém slovníku MorfFlex CZ 161115 (Hajič a Hlaváčová 2016). Do nástroje byla data vložena v podobě prostého textu. V lemmatizovaném textu nebyla pouze nahrazena slova svým základním tvarem, ale byly také odstraněny všechny pro daný slovník neznámé znaky, tedy i znaky z jiných abeced a emojies.

Po vložení lemmatizovaného textu zpět do prostředí RStudia, bylo nutné provést další úpravy textu, před aplikací tematického modelování. Veškerý text byl transformován na malá písmena a odstraněna byla také interpunkce a číslice. Součástí textu jsou také velmi frekventované výrazy jako spojky, zájmena a podobně, které by svým výskytem narušovaly výsledek modelu. Zároveň tato slova nenesou žádnou informaci o tématu, kterému se text

[1] MorphoDiTa: Morphological Dictionary and Tagger
[2] czech-morfflex-pdt-161115

věnuje. Proto byla tato stop slova[3] z textu odstraněna. Volba slov, která budou z textu vyřazena, je nezanedbatelnou součástí úpravy dat. Pozornost výzkumníka by měl mít při volbě stop slov kontext analyzovaných dat, kdy některá slova, jinak vhodná k odstranění mohou nést důležitý význam, nebo naopak mohou texty obsahovat slova, které jsou častá, ale vhodná k vyřazení (Lucas et al. 2015: 4). Z dat bylo například odstraněno slovo dnes, přestože není typickým stop slovem, které se v o aktuálním dění informujících facebookových příspěvcích vyskytuje velmi často, ale nenese pro analýzu důležitou informaci.

Příklad původního textu: *„Pomoc z Tchaj-wanu👍 Tchaj-wan nám posílá darem 25 plicních ventilátorů, které budou již příští týden rozděleny“…*

Příklad upraveného textu: *„pomoc tchajwan tchajwan posílat dar plicní ventilátor již příští týden rozdělit“…*

Úvodní zpracování a příprava textu je pro analýzu českého textu komplikovanější disciplínou, protože nástroje statistického zpracování textu jsou designovány pro práci v anglickém jazyce (Lucas et al. 2015: 3). Pro následné zpracování už bylo možné využívat standardní postupy, při využití knihoven *tm, openNLP, stm* nebo *LDAvis,* které nabízí vhodné nástroje (Nelson 2020: 10).

Pokud do analýzy vstupuje více datových korpusů, je možné předchozí kroky provádět na sloučených datech. Následující postup je pak aplikován na kormus každého jednotlivého autora facebookových příspěvků. U již rozdělených dat je vhodné vytvořit frekvenční tabulku výskytu slov. V případě rozdělení korpusu se mohou objevit slova, která se v textech systematicky vyskytují, do modelu ale nepřináší podstatnou informaci. V případě analyzovaných dat se jednalo zejména o jméno autora příspěvku, v případě ministryně Maláčové, nebo předsedy Senátu Vystrčila. Tato slova proto byla dodatečně odstraněna.

Data bylo dále nutné transformovat do podoby, s kterou umí pracovat knihovna *stm.* Tento krok byl proveden pomocí funkce *textProcessor* obsažené v knihovně *stm.* Pakliže by byl analyzovaný text v jazyce, jehož slovník je v této knihovně obsažen, bylo by možné všechny kroky popsané v této části provést pomocí výše zmíněné funkce. V případě, že jazyk není

[3] Kompletní seznam stop slov je uveden v příloze. Vychází ze seznamu, který použil Ligas (2020). Zde použitá stop slova jsou v základních tvarech, protože jsou aplikována na již lemmatizovaný text.

obsažen, je nutné využívat nástroje z jiných knihoven, které R obsahuje, externí nástroje jako je MorphoDiTa a v neposlední řadě také kreativitu výzkumníka, při nahrazování jednotlivých kroků zpracování textu.

Funkce *textProcessor* provádí také tokenizaci textu, tedy rozdělení souvislých textů příspěvků na jednotky analýzy. V případě vstupu pro *stm* je vhodné rozdělení na jednotlivá slova, kdy je tokenem právě slovo. Pomocí knihovny *quanteda* a její funkce *collocations* byly před použitím funkce *textProcessor* zohledněny některé časté spolu-výskyty slov a do analýzy vstoupily jako jeden token. Například dva tokeny „evropská unie“ byly transformovány na jeden token „evropská_unie“.

Před provedením samotného tematického modelování jsem vytvořil základní charakteristiku analyzovaných dat. V datovém korpusu čítajícím deset politiků přispívajících na svém profilu na Facebooku je celkem 5432 příspěvků. Mezi jednotlivými politiky jsou ale nezanedbatelné rozdíly. Počet příspěvků v roce 2020 v případě poslanců Petra Fialy a Tomia Okamury přesahuje 700, ale v případě senátora Jiřího Drahoše lehce přesahuje hranici 300 příspěvků a u poslance Zdeňka Vondráčka tohoto počtu ani nedosahuje. Celkový počet slov v příspěvcích jednotlivých autorů dosahuje 288 tisíc v případě příspěvků Okamury a plně tak ospravedlňuje využití STM, ale v případě dalších tří korpusů nedosahuje ani 15 tisíc slov, což je rozsah dat poměrně snadno obsažitelný lidským čtenářem.

Pro stabilitu modelu je kromě dostatečného množství textů důležitý také rozsah samotných textů. V ideálním případě mají texty přibližně stejný rozsah. Na základě srovnání průměrného počtu slov v textu, mediánu slov a maximálního počtu slov se je jako vhodná data jeví například příspěvky ministryně Maláčové, ministra Havlíčka nebo poslance Okamury. Všichni jmenovaní navíc publikovali vysoký počet příspěvků a malý počet z nich byl bez textu. Oproti tomu poslanec Volný nebo Filip byli v délce svých příspěvků v roce 2020 poměrně nekonzistentní. Značný počet jejich příspěvků pak neobsahoval žádný text a do analýzy tak nebude vstupovat. (tabulka 1)

tabulka 1 Základní charakteristika dat

Autor	Počet příspěvků	Počet slov v celém korpusu	Průměrná počet slov v příspěvku	Medián počtu slov v příspěvku	Maximální počet slov v příspěvku	Příspěvků bez textu
Drahoš	317	13903	45	38	284	7
Fiala	787	46346	60	44	700	1
Filip	399	39691	132	49	2616	97
Havlíček	650	33829	56	47	325	13
Maláčová	665	85271	133	107	739	6
Okamura	728	288429	403	240	3110	3
Vondráček	249	13705	57	42	663	2
Rakušan	629	57411	94	62	869	10
Volný	511	44445	101	36	4062	63
Vystrčil	497	14370	36	25	580	89

Pro samotné provedení STM je zásadní plně zpracovaný korpus textů. Při zpracování funkcí *textProcessor* byla slova, která se v korpusu vyskytla méně než dvakrát. Stejně tak texty které obsahovaly pouze jedno slovo, byly z analýzy vyřazeny. Tři z korpusů tak byly analyzovány v rozsahu nepřesahujícím 6 tisíc tokenů. (tabulka X)

tabulka 2 Charakteristika dat vstupujících do analýzy

Autor	Počet příspěvků	Analyzovaný počet příspěvků	Počet unikátních slov	Počet tokenů
Drahoš	317	308	1273	5525
Fiala	787	783	2378	19606
Filip	399	299	2698	14149
Havlíček	650	635	2186	15776
Maláčová	665	658	3778	32926
Okamura	728	724	8028	102826
Vondráček	249	247	1183	5187
Rakušan	629	617	3089	22114
Volný	511	424	2861	13843
Vystrčil	497	399	1187	5407

3.2 Aplikace tematického modelování

STM umožňuje do analýzy zahrnout proměnou, která ovlivňuje pravděpodobnost výskytu tématu v textu. Vzhledem k proměnlivosti témat byl jako tato kovariantní proměnná zvolen den v roce, transformovaný funkcí *s(),* obdobně jako je tomu s dokumentu představující STM (Roberts et al. 2019: 9).

Prvním krokem samotné analýzy bylo využití funkce *searchK*. Jak napovídá název samotné funkce, jejím úkolem je poskytnou orientační přehled možného počtu témat. Jak upozorňují autoři knihovny *stm,* tato analytická procedura neposkytuje odpověď na otázku, jaký je správný počet témat (Roberts et al. 2019: 12) Pro odhad vhodného počtu témat funkce vrací hodnoty zadržené pravděpodobnosti uvádějící změny modelu při modelování pouze na části dat (Wallach et al. 2009) a residua, udávající množství modelem nevysvětlených slov (Taddy 2011), které ukazují, zda model popisuje analyzovaná data, ale zároveň zda nedochází k přeučení[4] modelu. Modely s vhodným počtem témat dosahují vysokých hodnot zadržené pravděpodobnosti a naopak nízkých hodnot residuí (Silge 2018).

V rozsahu předpokládaného množství témat jsem vždy vytvořil několik modelů s rozdílnými počty témat. Cílem tohoto kroku bylo nalezení modelu, ve kterém jsou modelem identifikovaná témata smysluplně interpretovatelná. Tedy kdy je lidský čtenář schopen ztotožnit počítačem vytvořené téma se skutečným tématem, které se může vyskytovat v textech. K orientaci v obsahu identifikovaných témat je možné použít více nástrojů. Přímo pomocí knihovny *stm* je možné zobrazit nejčastěji se vyskytující slova v tématu, případně slova podle FREX hodnoty. Metrika slov s nejčastějším výskytem zobrazuje sémanticky koherentní témata, FREX pak hledá výskyt takových slov, která jsou sémanticky koherentní, ale zároveň specifická pro dané téma (Roberts et al. 2019: 10-11). Velmi intuitivním nástrojem je *LDAvis* (Sievert a Shirley 2014). Tato knihovna nabízí interaktivní vizualizaci modelu, umožňující prozkoumat výskyt jednotlivých slov v tématech a porovnat vzájemnou sémantickou podobnost témat. V této vizualizaci je také možné měnit parametr relevance slova a na témata tak nahlížet od optiky slov s nejčastějším výskytem, až po slova zcela specifická pro dané téma. Zejména tohoto nástroje jsem využívat při seznamování se s obsahem témat. Slova, která jsem takto identifikoval jako významná pro dané téma, také

[4] K přeučení dochází, pokud model příliš kopíruje analyzovaná data.

uvádím ve výsledcích analýzy. Vzhledem k obtížné přenositelnosti výpovědní hodnoty interaktivní vizualizace, nejsou podložena grafickým výstupem. Grafické výstupy uvedené v práci jsou tvořeny pomocí FREX slov.

3.3 Kvalitativní sonda

Tematické modelování umožňuje ve velkém objemu textů hledat témata, kterými jsou texty naplněny. Umožňuje tak určit základní vhled do dat, který by byl jinak pro rozsah dat obtížně dosažitelný. V každém textu navíc určí podíl jednotlivých témat. Výsledkem jsou texty, které téma obsahují a jsou pro dané téma podstatnými, ale i texty, ve kterých bylo téma identifikováno, ale na jeho podobě se podílí v minimální míře.

Pomocí funkce knihovny *stm findThoughts* je možné zobrazit ty texty příspěvků, které jsou pro konstrukci daného tématu v modelu nejdůležitější. Takovouto kvalitativní sondu jsem provedl do části analyzovaných dat. K tomu jsem využil otevřeného kódování. Vzhledem k metodě, která vede k výběru takových textů, jsem zvolil kódování spíše větších celků, jakými jsou v textu věty. Při kódování jednotlivých slov a slovních spojení by částečně docházelo k replikaci výsledků STM.

4 Praktická část

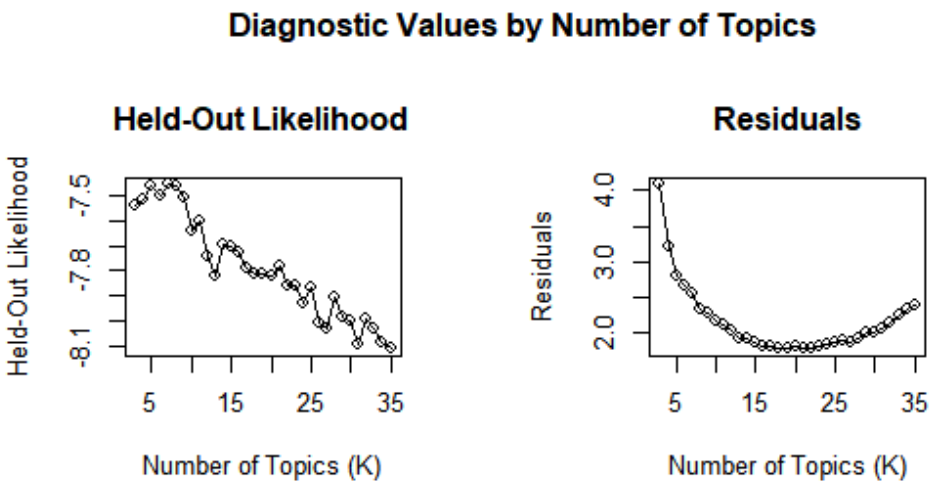
V této části přistoupím k samotné analýze dat. Nejprve za využití tematického modelu a následně v návaznosti na výsledky tohoto modelování i ke kvalitativní sondě do části dat.

4.1 Analýz latentních témat pomocí STM

4.1.1 Rakušan

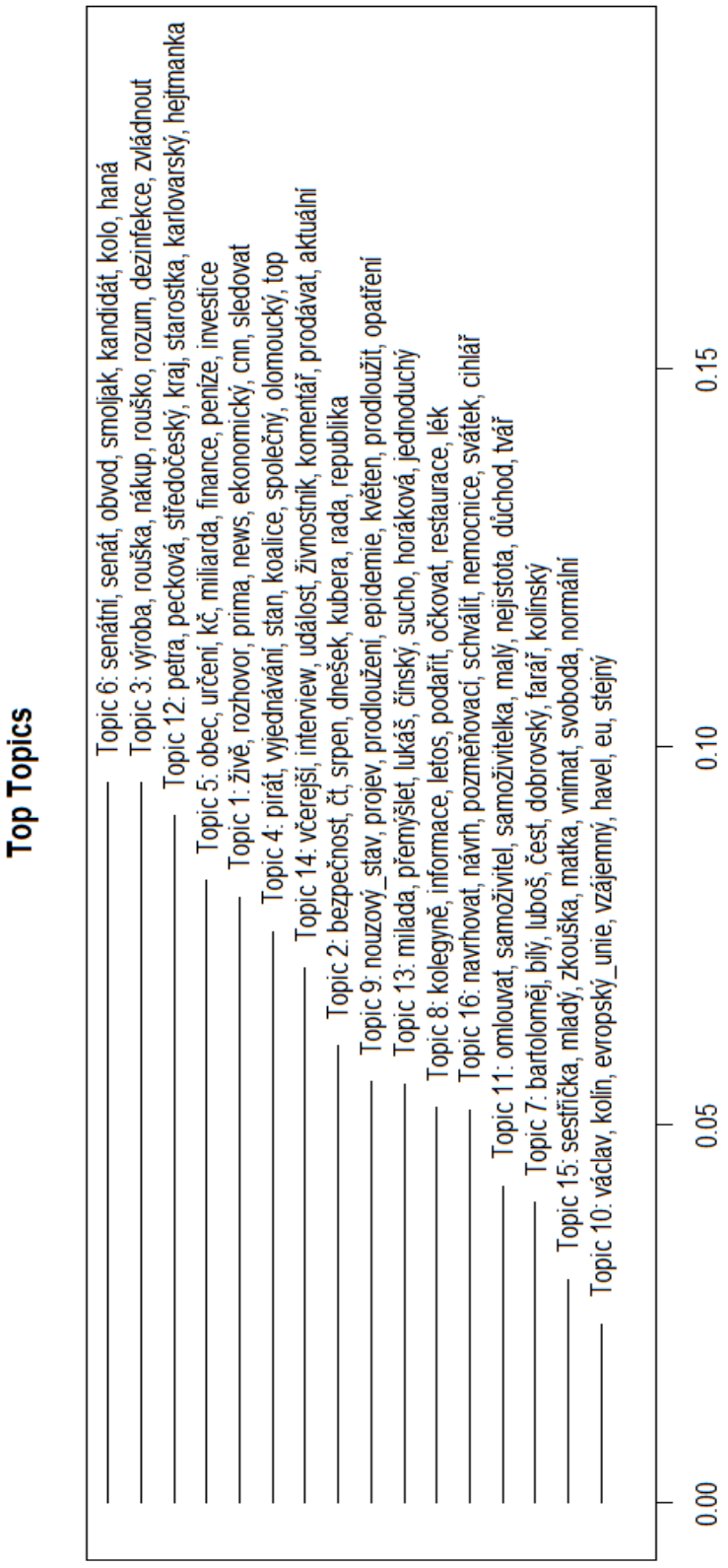
Pomocí funkce *searchK* a jí vrácených hodnot zadržené pravděpodobnosti a residuí byl vhodný počet témat odhadnut přibližně v rozmezí 13 až 20 témat. (graf 1)

graf 1 výsledek searchK s hodnotami zadržené pravděpodobnosti a residuí

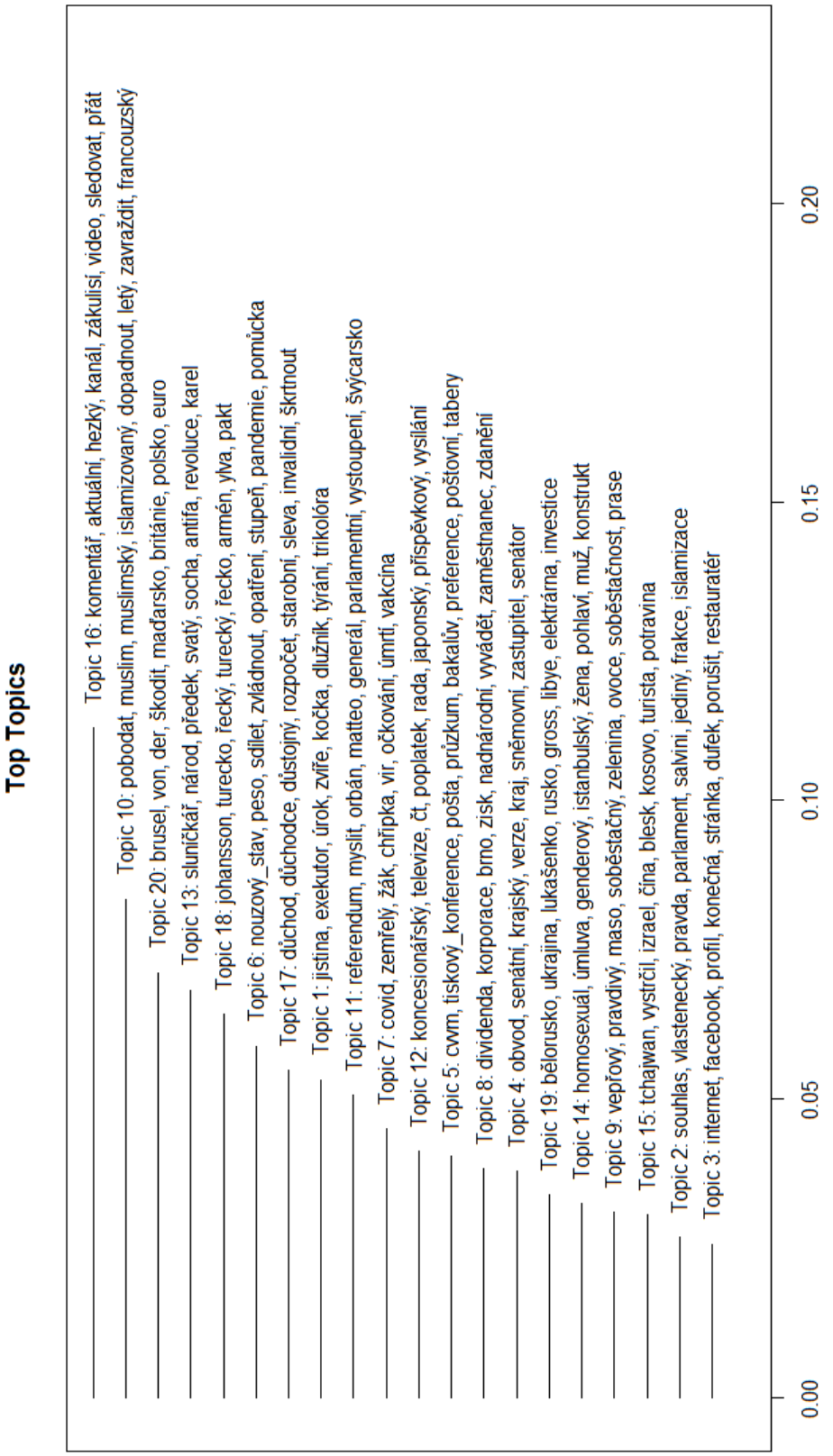


Modelu, který v příspěvcích poslance a předsedy STAN Víta Rakušana identifikoval 16 témat, uvádí jako výrazně zastoupené téma číslo 3. Můžeme v něm pozorovat důraz na *roušky, dezinfekci* ale také apel na *rozum*. Podle *LDAvis* jsou sémanticky blízká témata 9 a 8. Téma 9 se věnuje především nouzovému stavu nebo také protiepimickému systému (PES). Pro téma 8 jsou pak specifická slova *očkovat, očkování* nebo *vakcína* a tím se váže především k tématu vakcinace. (graf 2)

graf 2 Rakušan — identifikovaná latentní témata s relevantními slovy



graf 3 Okamura — identifikovaná latentní témata s relevantními slovy



4.1.2 Okamura

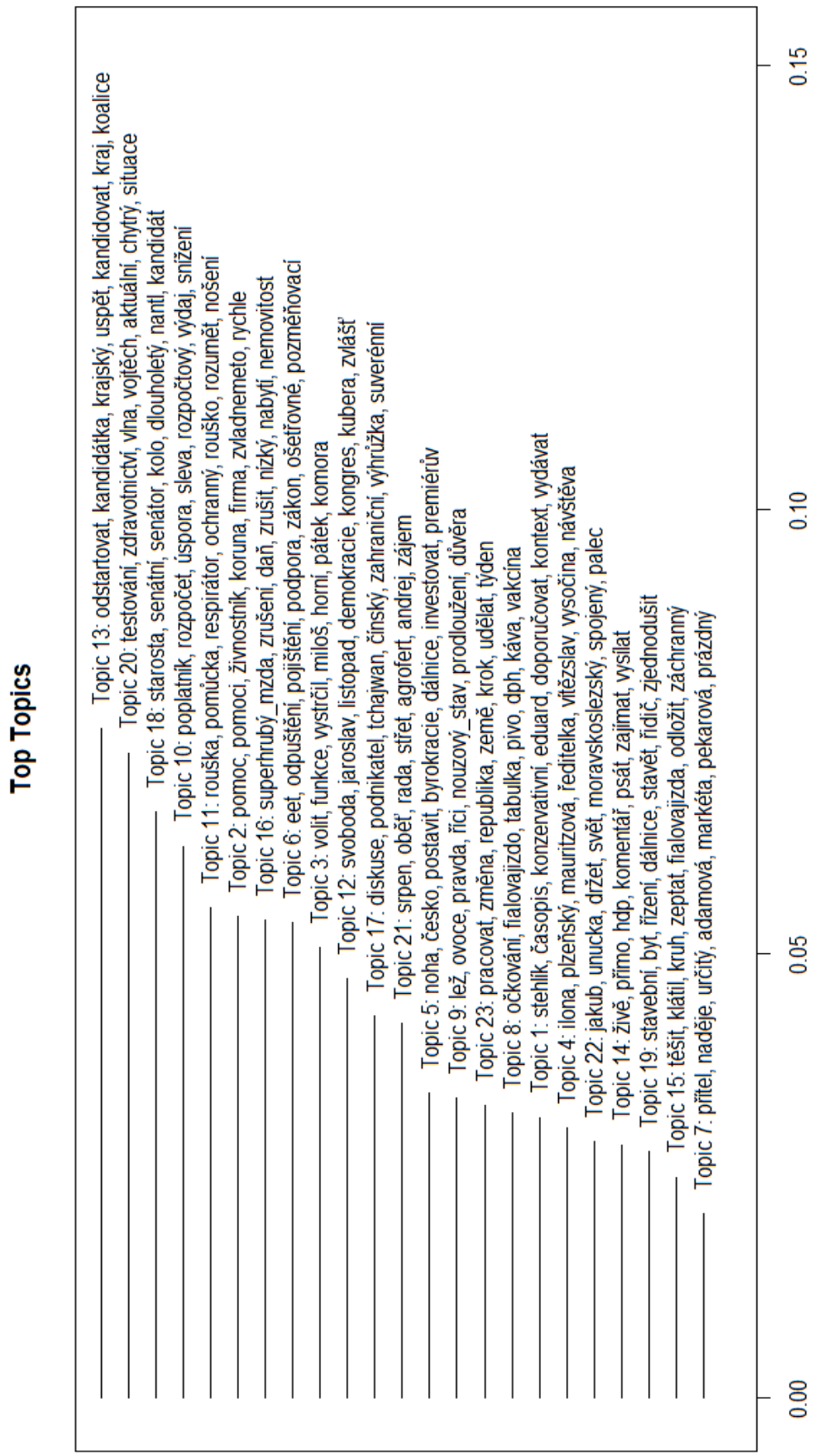
Na základě hodnot zadržené pravděpodobnosti a residuí, se zdá být vhodný počet témat spíše vyšší, mezi 15 a 30 tématy[5]. Vhodnou interpretaci nabízí model s 20 tématy. Téma koronaviru přímo zachycují témata 6 a 7. Téma 6 s charakteristickými slovy jako *nouzový_stav, pandemie* či *opatření* odkazuje k vlivu koronaviru na společnost. Oproti tomu téma číslo 7 podle slov v grafu *covid, chřipka, vir, očkování,* ale i dalších slov jedinečných pro toto téma, jako *onemocnění, testování* nebo *infekce,* odkazuje spíše na samotný koronavirus, vakcinaci a testování z medicínského pohledu. (graf 3)

4.1.3 Fiala

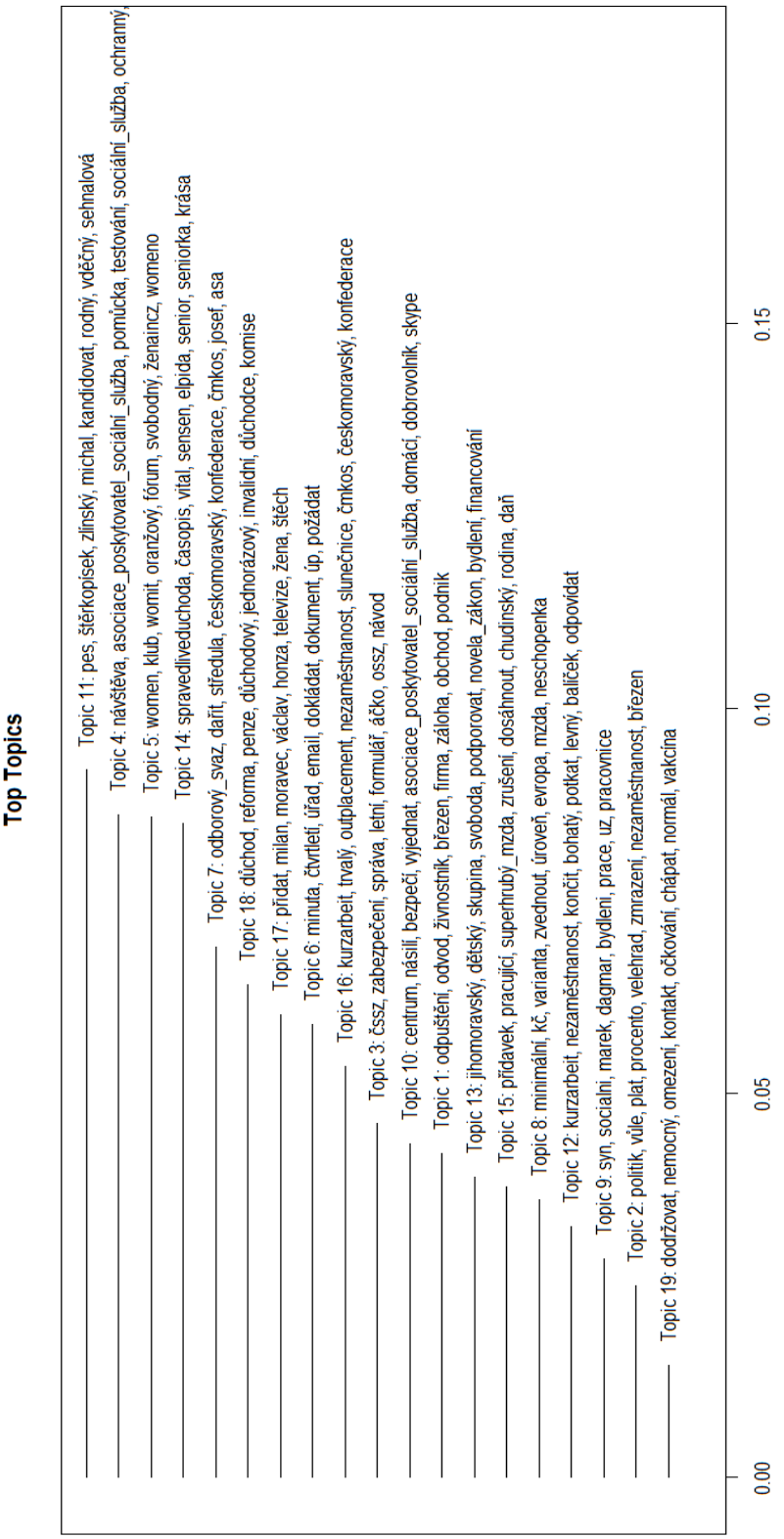
V případě příspěvků poslance a předsedy ODS Petra Fialy jsem použil model, který identifikoval 23 témat. Téma 20, které je v textech poměrně časté, popisuje zdravotnictví v souvislosti s koronavirem a vyskytují se v něm slova jako *testování, zdravotnictví, vlna* nebo bývalý ministr zdravotnictví Vojtěch. Téma číslo 11 pak poměrně specificky hovoří o *rouškách, respirátorech* a *ochranných pomůckách*. *Nouzový_stav* případně jeho *prodloužení* spíše v rovině politické jsou náplní tématu 9. (graf 4)

[5] Grafické výstupy funkce *searchK,* nejsou dále z důvodu jejich orientačního charakteru přiloženy. V některých případech tato funkce vrátila hodnoty, které se u jednoho z počtu témat lišily v hodnotách řádů. Graf tím byl zcela zkreslen. V takovém případě bylo nutné nahlédnout na vypočtené hodnoty jednotlivých ukazatelů.

graf 4 Fiala — identifikovaná latentní témata s relevantními slovy



graf 5 Maláčová — identifikovaná latentní témata s relevantními slovy



4.1.4 Maláčová

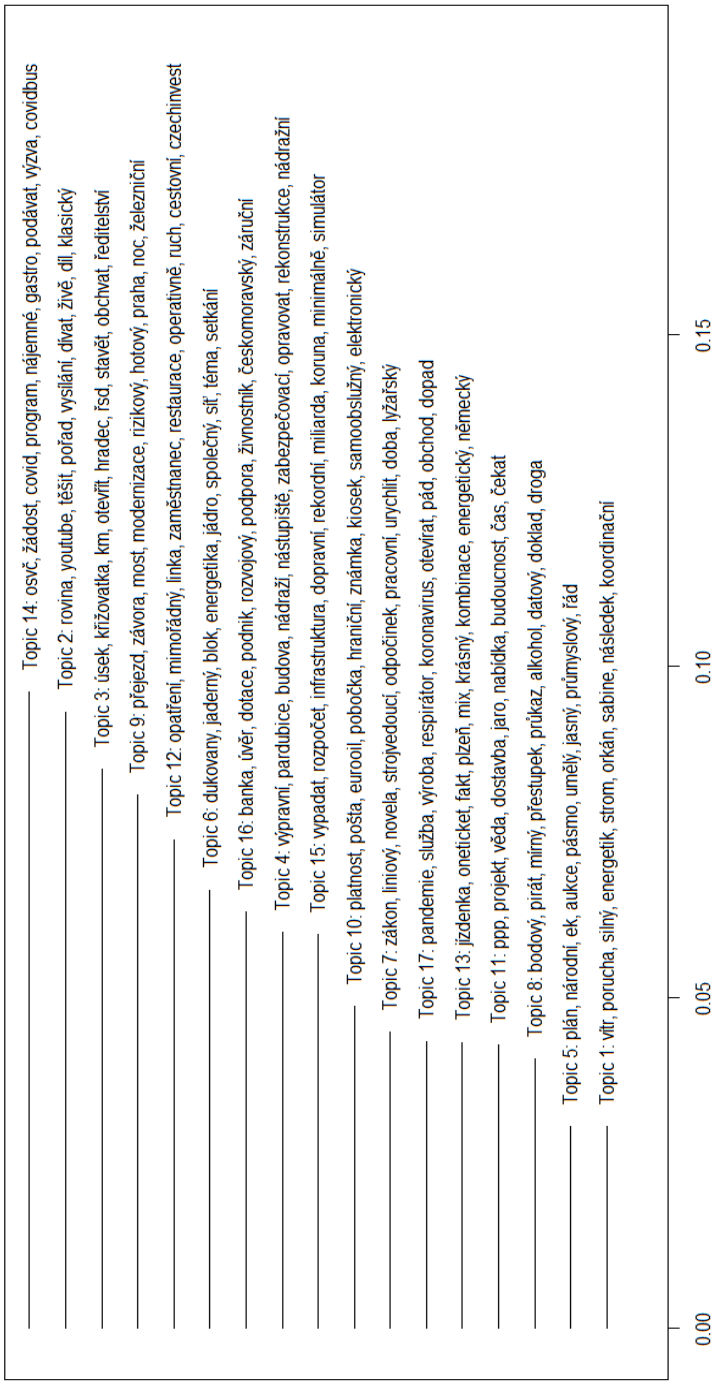
Z příspěvků ministryně práce a sociálních věcí Jany Maláčové byla jako stop slova dodatečně odstraněna slova *maláčová, jana, práce, sociální* a *ministerstvo.* Vzhledem k frekvenci výskytu by mohlo docházet ke zkreslení modelu.

Model, který v příspěvcích identifikoval 19 latentních témat, také v grafu, ale i skrze následné prozkoumaní pomocí *LDAvis* ukazuje, že téma koronaviru se projevuje v mnoha tématech. Téma koronaviru jako onemocnění, tak vykrystalizovalo až jako nejméně zastoupené téma číslo 19, reprezentované slovy *dodržování, nemocný, omezení, kontakt, očkování* nebo *vakcína.* Slova jako *koronavirus* nebo *testování* se vyskytují například v tématech 4 a 10, která se vztahují k sociálním službám, nebo k tématu 1, které kvůli výskytu slov *odvod, záloha* nebo *pojistník* můžeme ztotožnit s podporou podnikatelů a placením odvodů. (graf 5)

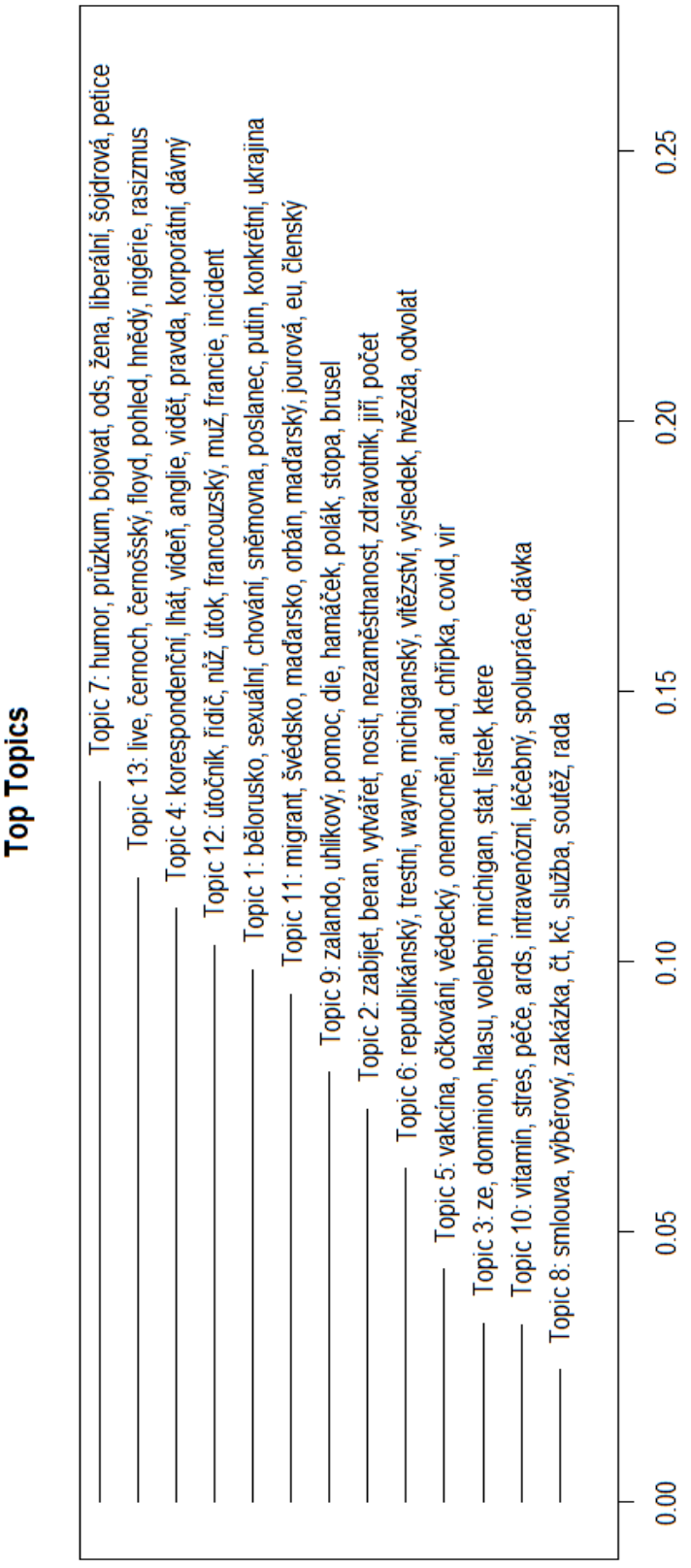
4.1.5 Havlíček

Strukturální, tematický model, pomocí kterého bylo v korpusu identifikováno 17 témat, odhalil jako téma s největším zastoupením v textu téma číslo 14. FREX slova v grafu u tohoto tématu spojují *osvč, žádost, covid, program.* Téma popisuje dopad koronaviru na podnikání osvč a nástroje ke zmírnění tohoto dopadu, které Karel Havlíček jako ministr průmyslu a obchodu nabízí. Na základě vizualizace *LDAvis* jsou si sémanticky blízká témata 12, 16, 17. Téma 12 pojednává o opatřeních proti šíření koronaviru, což spojuje s provozy restaurací a cestovním ruchem. Téma 16 se ke koronaviru váže především skrze dotační programy. Slova *covidnájemné,* nebo *covidkultura* jsou pro téma unikátní. Podobně jako v případě tématu 14 identifikuje především podporu podnikání, které bylo zavřené v důsledku vládních opatření. Posledním tématem z této skupiny, je téma číslo 17, které je typické slovy *respirátor, rouška* nebo dále *pandemie* či *epidemiolog.* Téma se zdá být nejvíce vztažené k samotné pandemii koronaviru. Ze čtyř popsaných témat je ale nejméně zastoupeno v textech. Koronavirus tak na facebookovém profilu ministra Havlíčka figuroval zejména prostřednictvím dopadu na podnikání, průmysl a ekonomiku. (graf 6)

graf 6 Havlíček — identifikovaná latentní témata s relevantními FREX slovy



graf 7 Volný — identifikovaná latentní témata s relevantními FREX slovy



4.1.6 Volný

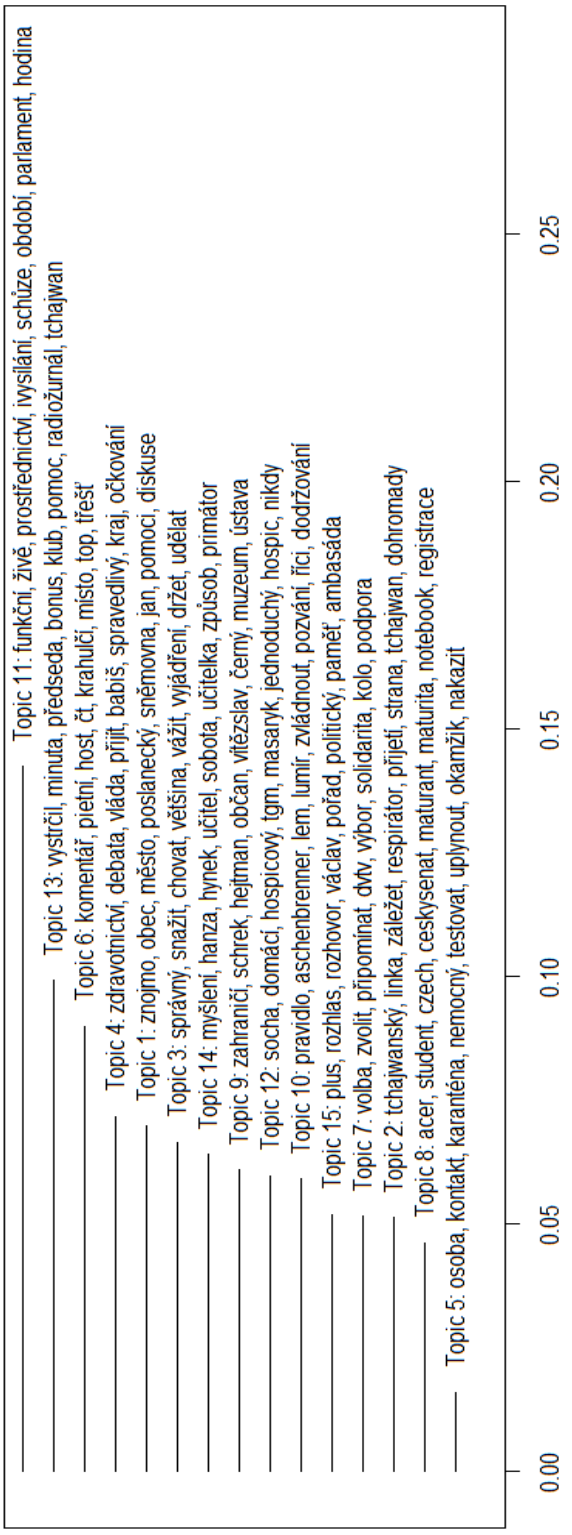
Z příspěvků poslance Lubomíra Volného byly z textů příspěvků dodatečně odstraněny stop slova *lubomír, volný* a *pairvolnystatuss.* Na takto upravených datech byl vytvořen model identifikující 12 témat. Přímé spojení tématu s koronavirem je z grafu 7 patrné v případě tématu 5, které slovy *vakcína, očkování, covid, vir* nebo *vědecký* přímo odkazuje k tématu koronaviru a vakcinace, a v případě tématu číslo 10, které je reprezentováno slovy *vitamín, léčebný* či *dávka* nebo *péče* a je možné je interpretovat s důrazem na prevenci proti onemocnění. Téma číslo 2 není reprezentováno tak zřejmými relevantními slovy, ale slova *zabíjet, zdravotník* nebo *beran,* což odkazuje na epidemiologa Jiřího Berana, definují vztah tématu ke koronaviru. Pro toto téma je také jedinečné slovo *isoprinosine,* odkazující na lék isoprinosine, který byl během pandemie diskutovaný v souvislosti s možnou léčbou onemocnění covid-19.

4.1.7 Vystrčil

Po analýze frekvence slov v textech byly z textů příspěvků předsedy Senátu Miloše Vystrčila odstraněny stop slova *miloš, vystrčil* a *předseda.* V takto upraveném textovém korpusu poskytuje sémanticky nejvíce koherentní témata model s 15 tématy zobrazený v grafu 8. V tomto modelu výrazně dominantní téma 11, které odkazuje na jednání senátu prostřednictvím slov *živě, ivysílání* nebo *schůze.* To napovídá, že komunikace na Facebooku ze strany Miloše Vystrčila je spíše informativního charakteru. Při využití *LDAvis* bylo možné odhalit, že slovo *covid,* se objevuje v tématu 12, které jej vztahuje k hospicové péči, a také v tématu 14, které i při pohledu do grafu obsahuje tématiku spojenou s *karanténou* a *testováním. Zdravotnictví* a *očkování* je také obsaženo v tématu 4 a můžeme v něm pozorovat rámování jako diskusi s vládou

Přestože v příspěvcích Miloše vystrčila bylo za pomoci STM možné identifikovat témata spojená s koronavirem, je i vzhledem k nedlouhým textům samotných příspěvků využití tematického modelování na hranici použitelnosti. Modelování nebylo schopné vyčlenit jednotlivá témata a při zvýšení jejich počtu docházelo ke snížení sémantické koherentnosti. Například téma 12 může být pro obraz pandemie koronaviru významné, ale je v něm obsažena i tématika sochy T.G.M.

graf 8 Vystrčil — identifikovaná latentní témata s relevantními FREX slovy



4.1.8 Filip

U příspěvků poslance a předsedy KSČM Vojtěcha Filipa hodnoty zadržené pravděpodobnosti a residuí vrácené funkcí *searchK* ukazují na vhodný počet témat mezi 8 a 20. Všechny vytvořené modely ale byly značně sémanticky nekonzistentní a téma přímo interpretovatelné jako popisující situaci způsobenou koronavirem nebylo možné identifikovat. Slova jako *pandemie, očkování* nebo *covid* se vyskytovala ve více tématech. Neschopnost vyčlenit v modelu samostatné téma koronaviru, je důsledkem malého rozsahu analyzovaných dat.

4.1.9 Drahoš a Vondráček

Vzhledem k rozsahu dat, která by do analýzy vstupovala v případě senátora Drahoše a předsedy Poslanecké sněmovny Vondráčka, a výsledkům analýzy příspěvků Vystrčila a Filipa, jsem od analýzy příspěvků posledních dvou politiků upustil.

4.2 Shrnutí tematického modelování

Výsledky tematického modelování ukazují, že témata přímo referující k pandemii koronaviru a jejím specifickým doprovodným aktérům, jako jsou roušky, očkování nebo nouzový stav, se v textech příspěvků politiků objevují rámována různě. V jednotlivých korpusech bylo možné identifikovat více než jedno takové téma. Téma zaměřené na roušky jsem určil v datech Rakušana, Fialy a Havlíčka. Dalším tématem, které se vyskytovalo u více politiků je nouzový stav, spojený s pandemií. Toto téma se jako samostatné vyskytuje u Rakušana, Okamury, a Fialy. S tématem vakcinace ve svých textech pracovali Rakušan, Okamura, Maláčová, Volný a Vystrčil. Testování na pozitivitu onemocnění covid-19 jsem jako samostatné téma identifikoval u Fialy a Vystrčila.

Na základě STM je patrné, že politici koronavirovou pandemii rámují různě. A jednotlivá témata se formovala na základě důrazu na ochranu dýchacích cest neboli roušky, pandemický nouzový stav, vakcinací nebo testováním na koronavirus (srov. Panda et al. 2020).

4.3 Kvalitativní sonda do textů

Na následujících dvou příkladech ukazuji, jaký je vztah mezi výsledky modelu a skutečným obsahem textů. Příspěvky Víta Rakušana jsem zvolil jako příklad mezi analyzovanými politiky typických dat. Příspěvky Tomia Okamury jsou v rámci dat výrazně rozsáhlejší, čímž by měly naplňovat požadavek STM, mají ale také nestabilní rozsah.

4.3.1 Rakušan – roušky, nouzový stav a očkování

Na základě kvalitativní sondy do devíti nejvýznamnějších příspěvků, ve kterých byla identifikována tři témata spojená s koronavirem, tedy tří příspěvků charakteristických pro každé ze tří témat, jsem dospěl k následujícím zjištěním.

Dva z příspěvků, ve kterých je významně zastoupeno téma 3, které jsem spojil s rouškami, jsou příspěvky, které otevírají více témat. Autor se v nich systematicky v odrážkách vyjadřuje k většímu množství témat. V souvislosti s rouškami je v textech významný aspekt aktivity jednotlivce konstruovaný prostřednictvím šití roušek pro sebe i ostatní nebo pomoci seniorům.

V příspěvcích významných pro téma nouzového stavu je patrný apel na návrat k *normálnímu životu,* který se střetává kroky vlády v souvislosti s opatřeními. Cesta k normálnímu životu ale vede skrze právě skrze protiepidemická opatření. *„ Vypněme mentalitu ‚nouzového stavu' a nastavme se na svobodu a zodpovědnost."* V komunikaci tohoto tématu můžeme skrze důraz na zodpovědnost spatřovat snahu o internalizaci pravidel, což umožní již zmiňovaný *normální život*.

Texty, které jsou nejvýznamnější pro téma, které jsem na základě STM charakterizoval jako téma, jehož hlavní náplní je očkování, toto téma prakticky neartikulují. V textech je ale podstatný motiv velkého očekávání a naděje na zlepšení situace. Jedním z významných slov pro téma z STM je *podařit.* Tato očekávání a neděje se projevují ve vztahu k rozvolňování. „*Až se ve čtvrtek otevřou naše malé obchody, jděme do nich nakupovat*." Ale i naopak k zavedeným opatřením. „*Lidé musí věřit tomu, že jejich oběť a omezení běžných životních návyků má smysl."* Očekávání je vkládáno právě i do očkování, kdy „*bez očkování se tady v téhle mizérii budeme plácat ještě minimálně rok*."

4.3.2 Okamura – chřipka a nouzový stav

Dalším vybraným politikem v této části analýzy je Tomio Okamura. Jako reprezentativní texty byly zvoleny tři příspěvky, v kterých bylo téma identifikováno. Celkem tedy šest příspěvků. Pomocí STM jsem určil dva základní rámce, kterými bylo téma koronaviru v příspěvcích Tomia Okamury rámováno. Rámec koronaviru jako nemoci se ve výstupu STM vyskytovalo slovo žák. Při bližším pohledu do dat, je zjevné, že toto spojení je následkem otevírání tématu školství a koronaviru ve stejných příspěvcích. Dva z příspěvků identifikovaných jako klíčové pro dané téma obsahovali i značnou část textu věnující se školství. Nabízela by se tak možnost vrátit se k analýze modelováním a hledat počet témat, kde se tato témata oddělí. Při zvyšování počtu témat ale docházelo i k rozpadu části tématu spojené s koronavirem.

Ve vztahu k tématu koronaviru je kladen důraz zejména na odbornost, a to především tu medicínskou. Toto zjištění tak podporuje interpretaci tohoto tématu. Tento aspekt naplňuje především přepis proslovu z poslanecké sněmovny, ve kterém je velké míře argumentováno statistickými daty i odkazy na odborné autority. Věta, „*dovolil bych si citovat imunology,*" tak vystihuje rámování koronaviru jako medicínského fenoménu.

Rámec, který ke koronaviru přistupoval jako ke společenskému problému, je v jednom případě reprezentován přepisem projevu z Poslanecké sněmovny, tedy rozsáhlejším textem. Je vyzdvihován zejména obava, že *krize zdravotní přejde do krize ekonomické.* Dále obava z následků na život jednotlivce v podobě nemožnosti splácet hypotéku nebo nebezpečí exekuce jsou v textech přítomny a ukazují, že téma koronaviru a jeho dopadů na společnost je spjato s běžnou sociální problematikou.

4.4 Diskuse

4.4.1 Výsledky analýzy

Představy vyvozené o obsahu příspěvků politika na základě témat, která daný politik komunikoval, i jak rámoval onemocnění covid-19 a s ním spojené doprovodné jevy, byly po provedení kvalitativní analýzy značně obohaceny.

Z výsledků kvalitativní analýzy příspěvků Víta Rakušana, je patrná hodnotová rovina aktivity jednotlivce a sounáležitosti, přijetí a internalizaci zodpovědného chování nebo vkládání naděje do změn, které probíhají. Až kvalitativní analýza ukázala, jaký význam je jednotlivým tématům dáván. V případě příspěvků Tomia Okamury pak tento rozdíl není tak významný. Že je na téma koronaviru, jako medicínského fenoménu touto optikou i nahlíženo, byl předpoklad z výsledků tematického modelu. Apel na odbornost nebo častá argumentace statistickými daty, je pak zřejmá až při čtení textů, a to i proto, že čísla byla z dat pro modelování odstraněna. V komunikaci ale čísla mají nezastupitelnou roli a vytvářejí argument sdělení.

4.4.2 Aplikace metody

Strukturální tematické modelování se v prvním kroku analýzy ověřilo jako vhodná metoda pro analýzu facebookových příspěvků. Na základě jeho výsledků bylo možné identifikovat témata v kterých se tématika koronaviru vyskytovala, a zmapovat tak rámování tohoto fenoménu v roce 2020.

Z povahy této metody, poskytovaly výsledky kvantitativní analýzy více rozsáhlých dat výstupy, jejichž interpretaci nepovažuji za obtížnou. V případě, že model pracoval s menším množstvím textů, stával se výstup obtížněji interpretovatelným. I v případě dat, která bychom mohli označit jako nevhodná, bylo možné v případě příspěvků Miloše Vystrčila, téma koronaviru identifikovat a navrhnout interpretační rámec.

V případě metodologického postupu se následná kvalitativní analýza textů, v kterých je téma, které je předmětem zájmu, nejvíce zastoupeno se jeví jako problematická. Na datech, která jsou pro STM ideální stabilním rozsahem textů a v případě této aplikace rozsahem v nižších stovkách slov, v této práci demonstrované příspěvky Víta Rakušana, tato metoda může vykazovat dobré výsledky. V případě analýzy rozsahem méně stabilních dat, kterými jsou v této práci příspěvky Tomia Okamury, může STM pomoci odhalit latentní témata, která se v textech vyskytují, následné využití této metody pro třídění kvalitativní analýzy ale může být značně problematické, jak ukazuje příklad příspěvků Tomia Okamury.

4.4.3 Alternativy metodologie

Alternativní metodou by byla analýza témat pomocí tematického modelování, která by poskytla přehled o tématech v textech, následovaná výběrem textů pro kvalitativní analýzu pomocí prostého výskytu klíčových slov v textu. Na základě tohoto postupu by představa, kterou si výzkumník o daném tématu vytvoří při intepretaci klíčových slov, byla přenesena přímo na texty. Byl by tak omezen vliv slov, která mohou být v tématu velmi zřídka zastoupena, přesto jsou pro něj unikátní. V případě velkého množství textů je ale pomocí v této práci požitého postupu velmi pravděpodobné, nalézt jako pro téma nejvýznamnější právě takové texty, ve kterých jej reprezentují v podobě blízké interpretaci modelu.

Další možností, jak zvýšit efektivitu toho postupu je rozdělení rozsáhlých textů. Takový postup nachází opodstatnění při bližším pohledu na data použitá v této práci. Jak zmiňuji výše, byly některé delší texty, které vstoupily do kvalitativní analýzy poměrně jasně členěny. Tato skutečnost napovídá, že texty je tak možné rozdělit na části, podle členění na odstavce a zachovat celistvost jednotlivých tematických celků. Vzhledem k tomu, že takový postup by vycházel z logického členění samotným autorem textu, mohl by přispět k výsledkům modelu, které by více sledovaly autorovo chápání tematického prostoru. V takovém případě by bylo vhodné využít algoritmus, uzpůsobený k analýze krátkých textů.

Zlepšení výsledů samotného modelování by mohla přinést volba algoritmu, který by lépe odpovídal požadavkům dat i výzkumnému cíli. Vzhledem k množství modelů je pravděpodobné, že při konfrontaci s konkrétním výzkumným cílem bude existovat možnost vhodnějšího modelu, než je STM. Tyto modely ale nenabízí tak robustní a dobře integrovanou paletu nástrojů jako je například ta dostupná v knihovně *stm.* Jejich chování při aplikaci na konkrétní data pak není podpořeno literaturou, jako je tomu v případě STM.

4.4.4 Politici sedí na více židlích zároveň

Facebook je ze strany politiků v českém prostředí hojně využívaným komunikačním kanálem. Není ovšem zdaleka jediným sociálním médiem, prostřednictvím kterého politici komunikují. Prostor se proto tak vzniká pro výzkum, které by analyzoval všechny platformy, které politik využívá. Takový výzkumný design má potenciál, zachytit různé způsoby komunikace, které jsou podmíněny samotnými sociálními sítěmi i rozdíly v komunikaci

v důsledku jiných konzumentů informací. Zatímco na Facebooku se délka textů může významně lišit, jak jsem demonstroval v této práci, je délka příspěvků na Twitteru omezena. V případě analýzy dat z obou sociálních sítí se tak výzkumník musí vypořádat rozdílnou povahou dat.

Uvážíme-li jakou popularitu získal během pandemie koronaviru instagramový profil poslance Dominika Feriho, na kterém se zněkolikanásobil počet sledujících a v současnosti přesahuje milion sledujících, je tento komunikační kanál nepochybně klíčový pro to, jak tento politik komunikuje téma koronaviru. Začlenit do výzkumu tematického rámování koronaviru ze strany tohoto politika Instagram, by tak jistě bylo žádoucí, ale také výzvou pro výzkumný design.

4.4.5 Víc než jen text

Jak zmiňuji v kapitole zabývající se úpravou dat, z textů všech příspěvků byly odstraněny tzv. emojies. Tento komunikační nástroj není v příspěvcích politiků na Facebooku výjimkou. Analyzovaná data zahrnovala značné množství těchto významově i vizuálně specifických znaků emojies v elektronické komunikaci sehrávají významnou roli a to nejen pro příjemce informací (Kaye et al. 2017). Emojies mohou plnit funkci syntaktickou, která ovšem pro tematické modelování, jakožto metodu pracující s texty jako s balíky slov, není významná. Emojies mají ale především funkci sémantickou, v textu vytvářejí konotace a vyjadřují emoce (Arafah a Muhammad 2019).

Zejména emojies, jako nástroj pro vyjádření emocí by měl pro metody tematického modelování být důležitý prvek komunikace. Pokud autor ať už je jím politik, či nikoliv využívá emojies pro vyjadřování emocí, nahrazuje tímto symbolem slovo, nebo slova, která by jinak vstoupila do modelu a mohla by se podílet na vytváření témat. Nemusí se nutně jednat jen o nástroj k vyjádření emocí. Součástí unicódu jsou také emojies dopravních prostředků, zvířat nebo vlajky. V příspěvcích Tomia Okamury se například vyskytují série emojies české národní vlajky. Takový komunikační prvek nese značnou informační hodnotu.

Jako běžné znaky inicódu mohou být emojies tokenizovány a do analýzy vstupovat stejně jako jednotlivá slova textu. Tento prostý krok se ale může potýkat při provádění analýzy za použití standardních nástrojů a postupů. Knihovny a funkce, které jsem používal při zpracování dat, jsou designovány pro práci s anglickým textem a již cizí jazyky komplikují

úpravu textu. Pro výzkumníka, který má s programovacím jazykem zkušenosti, nebude taková úprava problémem. Z pohledu popularizace tematického modelování a nárůstu využívání této metody k analýze textů se ale může jednat o bariéru.

Dalším aspektem je informace, která se neskrývá v samotném textu. Typický příspěvek na Facebooku se neskládá pouze z textu. Jak ukazují analyzovaná data, ze všech 5432 příspěvků obsahovala více než polovina přiložený obrázek. U 1255 příspěvků pak bylo jeho součástí video. Informace, kterou nesou tato média, tak do analýzy nevstupuje vůbec.

5 Závěr

V této bakalářské práci diskutuji využití tematického modelování, konkrétně strukturálního tematického modelu při využití tematické analýzy na datech sestávajících z facebookových příspěvků politiků v rozsahu jednoho roku. Dále se tato práce zabývá využitím tematického modelování jako první fáze výzkumu s následnou kvalitativní analýzou textů. V textech příspěvků jednotlivých politiků byly tímto postupem odhaleny podoby, které dávají politici svými příspěvky na sociálních sítích tematickému základu koronaviru a jak tento fenomén tematicky rámují. Tato práce nepřináší definitivní odpověď na otázku, zda je pomocí tematického modelování vhodné provádět výběr textů pro kvalitativní analýzu v paradigmatu mixed method. Výsledek analýzy ale ukazuje, že tento postup může při použití metodologického rámce na vhodných datech vést k volbě textů vhodných pro kvalitativní analýzu.

6 Summary

This bachelor thesis discusses an application of structural topic model on the data of facebook posts of politicians from a year long period of the year 2020 which was dramaticly affected by the pandemic of coronavirus. This thesis further proposes a mixed method framework consisted of follow up qualitative analysis of for the topic important posts. This followup was succesfully performed on a sample of data, providing an insight view over topic framing of covid-19 during the pandemic. Based on the aplication of method, this thesis do not give a clear answer whether topic modeling is a fitting tool for topic selection of

significant posts. Resuts of the analysis show the possibility application of the framework when used on suitable data.

7 Použitá literatura a zdroje

ARAFAH, Burhanuddin a Hasyim MUHAMMAD, 2019. Linguistic functions of emoji in social media communication. **35**, 558–574.

BARRIOS, John M. a Yael HOCHBERG, 2020. *Risk Perception Through the Lens of Politics in the Time of the COVID-19 Pandemic*. w27008. B.m.: National Bureau of Economic Research.

WEGERIF, Rupert; MERCER, Neil. Using computer-based text analysis to integrate qualitative and quantitative methods in research on collaborative learning. Language and Education, 1997, 11.4: 271-286.

BLEI, David M., 2012. Probabilistic Topic Models. *Commun. ACM*. **55**(4), 77–84. ISSN 0001-0782.

BLEI, David M., Andrew Y. NG a Michael I. JORDAN, 2002. Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems 14*. B.m.: MIT Press, s. 601–608.

BLEI, David M., Andrew Y. NG a Michael I. JORDAN, 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. **3**, 993–1022. ISSN 1533-7928.

BOLDEN, Richard a Jean MOSCAROLA, 2000. Bridging the Quantitative-Qualitative Divide: The Lexical Approach to Textual Data Analysis. *Social Science Computer Review*. **18**(4), 450–460. ISSN 0894-4393.

CRAIN, Steven P., Ke ZHOU, Shuang-Hong YANG a Hongyuan ZHA, 2012. Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond. In: *Mining Text Data*. Boston, MA: Springer US, 129–161. ISBN 978-1-4614-3223-4.

EDELMANN, Achim, Tom WOLFF, Danielle MONTAGNE a Christopher BAIL, 2020. Computational Social Science and Sociology. *Annual Review of Sociology*. **46**. 61-81.

HÁJEK, Martin, 2014. *Čtenář a stroj: vybrané metody sociálněvědní analýzy textů*. Vyd. 1. Praha: Sociologické nakladatelství (SLON). ISBN 978-80-7419-161-9.

HAJIČ, Jan a Jaroslava HLAVÁČOVÁ, 2016. MorfFlex CZ 161115. *http://ufal.mff.cuni.cz/morfflex* [online]. [vid. 2021-05-02]. Dostupné z: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1834

CHAKRABARTI, Parijat a Margaret FRYE, 2017. A mixed-methods framework for analyzing text data: Integrating computational techniques with qualitative methods in demography. *Demographic Research*. **37**, 1351–1382. ISSN 1435-9871.

CHENG, Xian, Qiang CAO a Stephen Shaoyi LIAO, 2020. An overview of literature on COVID-19, MERS and SARS: Using text mining and latent Dirichlet allocation. *Journal of Information Science.* ISSN 0165-5515.

JELODAR, Hamed, et al, 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*. **78**(11), 15169–15211. ISSN 1573-7721.

JÜNGER, Jakob a Till KEYLING, 2019. *Facepager. An application for generic data retrieval through APIs.* [online]. Dostupné z: https://github.com/strohne/Facepager

KAPLAN, Andreas M. a Michael HAENLEIN, 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*. **53**(1), 59–68. ISSN 0007-6813.

KAYE, Linda K., Stephanie A. MALONE a Helen J. WALL, 2017. Emojis: Insights, Affordances, and Possibilities for Psychological Science. *Trends in Cognitive Sciences*. **21**(2), 66–68. ISSN 1364-6613.

KRONICK, JANE C., 1997. Alternativní metodologie pro analýzu kvalitativních dat / Alternative Methodologies for the Analysis of Qualitative Data. *Sociologický Časopis / Czech Sociological Review*. **33**(1), 57–67. ISSN 0038-0288.

L. LI, Y. SUN, a C. WANG, 2018. Semantic Augmented Topic Model over Short Text. In: *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*: *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*. s. 652–656.

LIGAS, Aleš, 2020. Jak jazyk ovládá mysl: Obsahová kvantitativní analýza rétoriky hnutí ANO, SPD a IvČRN ve vztahu k totalitnímu jazyku Třetí říše

LIMAYE, Rupali Jayant, et al, 2020. Building trust while influencing online COVID-19 content in the social media world. *The Lancet Digital Health*. **2**(6), e277–e278. ISSN 2589-7500.

LIU, Yan, Alexandru NICULESCU-MIZIL a Wojciech GRYC, 2009. Topic-link LDA: joint models of topic and author community. In: *the 26th Annual International Conference*: *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. Montreal, Quebec, Canada: ACM Press, s. 1–8. ISBN 978-1-60558-516-1.

LUCAS, Christopher, et al. 2015. Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*. **23**(2), 254–277. ISSN 1047-1987.

MACEK, Jakub, 2012. Nevyčíslitelné porozumění: kvalitativní výzkum online sociálních sítí. *ProInflow*. **4**(1). ISSN 1804-2406.

MACKOVÁ, Alena, 2018. *Nová média v politické komunikaci: politici, občané a online sociální sítě*. Brno: Masarykova univerzita. Ediční řada Monografie. ISBN 978-80-210-8745-3.

MIRBABAIE, Milad, et al. 2020. Social media in times of crisis: Learning from Hurricane Harvey for the coronavirus disease 2019 pandemic response. *Journal of Information Technology*. **35**(3), 195–213. ISSN 0268-3962.

MURDOCH, David, et al. 2020. Politicians: please work together to minimise the spread of COVID-19. *The New Zealand Medical Journal*. **133**(1511), 7–8.

NELSON, Laura K., 2017. lknelson/computational-grounded-theory. *GitHub* [online] [vid. 2021-03-23]. Dostupné z: https://github.com/lknelson/computational-grounded-theory/data/README.txt

NELSON, Laura K., 2020. Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*. **49**(1), 3–42. ISSN 0049-1241.

NGUYEN, Dat Quoc, et al. 2015. Improving Topic Models with Latent Feature Word Representations. *Transactions of the Association for Computational Linguistics; 3, 299-313*

PALA, Karel, et al. 2010. Český WordNet 1.9 PDT.

PANDA, Anmol, Divya SIDDARTH a Joyojeet PAL, 2020. COVID, BLM, and the polarization of US politicians on Twitter. *arXiv:2008.03263*

Q. LIQING, J. WEI, L. HAIYAN, a F. XIN, 2019. Microblog Hot Topics Detection Based on VSM and HMBTM Model Fusion. *IEEE Access*. **7**, 120273–120281. ISSN 2169-3536.

QIANG, Jipeng, et al. 2018. *STTM: A Tool for Short Text Topic Modeling*.

ROBERTS, Margaret E., Brandon M. STEWART a Edoardo M. AIROLDI, 2016. A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association* . **111**(515), 988–1003. ISSN 0162-1459.

ROBERTS, Margaret E., Brandon M. STEWART a Dustin TINGLEY, 2019. stm: An R Package for Structural Topic Models. *Journal of Statistical Software*. **91**(2), 1–40. ISSN 1548-7660.

ROBERTS, Margaret E, et al. 2013. The Structural Topic Model and Applied Social Science. In: *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*. s. 1-20.

RODRIGUEZ, Maria Y. a Heather STORER, 2020. A computational social science perspective on qualitative data exploration: Using topic models for the descriptive analysis of social media data*. *Journal of Technology in Human Services*. **38**(1), 54–86. ISSN 1522-8835.

RYOO, Joseph a Neil BENDLE, 2017. Understanding the Social Media Strategies of U.S. Primary Candidates. *Journal of Political Marketing*. **16**(3–4), 244–266. ISSN 1537-7857.

SIEVERT, Carson a Kenneth SHIRLEY, 2014. LDAvis: A method for visualizing and interpreting topics. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, Maryland, USA: Association for Computational Linguistics, s. 63–70.

SILGE, Julia, 2018. Training, evaluating, and interpreting topic models. *Julia Silge* [online] [vid. 2021-04-19]. Dostupné z: https://juliasilge.com/blog/evaluating-stm/

SNELSON, Chareen L., 2016. Qualitative and Mixed Methods Social Media Research: A Review of the Literature. *International Journal of Qualitative Methods*. **15**(1).

STIEGLITZ, Stefan a Linh DANG-XUAN, 2013. Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining*. **3**(4), 1277–1291. ISSN 1869-5469.

STIER, Sebastian, Lisa POSCH, Arnim BLEIER a Markus STROHMAIER, 2017. When populists become popular: comparing Facebook use by the right-wing movement Pegida and German political parties. *Information, Communication & Society*. **20**(9), 1365–1388.

TADDY, Matthew A., 2012. On Estimation and Selection for Topic Models. In: *Artificial Intelligence and Statistics*. PMLR, p. 1184-1193.

WALLACH, Hanna M., 2006. Topic Modeling: Beyond Bag-of-Words. In: *Proceedings of the 23rd International Conference on Machine Learning*. s. 977–984. ISBN 1-59593-383-2.

WALLACH, Hanna M., et al. 2009. Evaluation methods for topic models. In: *the 26th Annual International Conference*: *Proceedings of the 26th Annual International Conference on Machine Learning.* s. 1–8.

WANG, Yi-Chia, Moira BURKE a Robert E. KRAUT, 2013. Gender, Topic, and Audience Response: An Analysis of User-Generated Content on Facebook. In: *Proceedings of the*

SIGCHI Conference on Human Factors in Computing Systems. s. 31–34. ISBN 978-1-4503-1899-0.

X. CHENG, X. YAN, Y. LAN, a J. GUO, 2014. BTM: Topic Modeling over Short Texts. *IEEE Transactions on Knowledge and Data Engineering*. **26**(12), 2928–2941. ISSN 2326-3865.

X. WU a C. LI, 2019. Short Text Topic Modeling with Flexible Word Patterns. In: *2019 International Joint Conference on Neural Networks (IJCNN)*: *2019 International Joint Conference on Neural Networks (IJCNN)*. s. 1–7. ISBN 2161-4393.

YUM, Seungil, 2020. Social Network Analysis for Coronavirus (COVID-19) in the United States. *Social Science Quarterly*. **101**(4), 1642–1647. ISSN 1540-6237.

Institut sociologických studií, FSV UK

Projekt bakalářské práce

Předpokládaný název práce

Mixed method: analýza krátkých biografických výpovědí

Námět práce zahrnující formulaci a vstupní diskusi poznávacího problému

Sociologický výzkum je založen na rozmanitých metodách sběru dat a analýzy. Jejich vhodnou kombinací se snaží vyžívat silných stránek metodologických postupů tak, aby co nejpřesněji zodpověděl otázky sociálního světa. Dvě základní paradigmata, kvalitativní a kvantitativní výzkum se v současnosti mísí v podobě *mixed method* výzkumného designu.

Svoji bakalářskou práci pojmu jako práci metodologickou, zkoumající jednu z možných podob *mixed method* designu. Tato podoba výzkumného designu, kterou blíže popíši v následující části si klade za cíl poskytnout výstup blížící se výstupu kvalitativního výzkumu. Tento *mixed method* design proto budu srovnávat s výstupem klasického kvalitativního šetření založeného na analýze hloubkových rozhovorů. Tématem výzkumu bude alkohol v životě mládeže. Chtěl bych se opřít zejména o výzkum z roku 2016 provedený Sociologickým ústavem AV ČR. Nepředpokládám významné změny v postoji mládeže k alkoholu od roku 2016 a proto budou výsledky obou výzkumných designů srovnatelné. Téma svojí práce však zúžím na opilost, která je s konzumací alkoholu spojená. Konkrétní výzkumné otázka praktické části se bude ptát, jakými způsoby mluví mladí lidé o opilosti a také, co pro ně opilost znamená.

Předpokládané metody zpracování

Za použití dotazníků založeného na otevřený otázkách, ve který respondent stručně popíše, co pro něj znamená opilost a v podobě příběhu zrekonstruuje svoji poslední zkušenost s opilostí, je mým cílem získat zejména krátký narativ. Cílem je přimět respondenty, aby

výpovědi založili i na hodnocení, ne pouze prostém popisu. Výpověď proto stylizují do podoby, jako kdyby danou skutečnost líčili kamarádovi. Popis zkušenosti s opilostí bude mít minimální rozsah 150 slov, aby byl vhodný pro analýzu. Doplnění o malé množství identifikačních otázek, poskytne další rozměry pro další analýzu.

Takto sesbírané odpovědi, budou deduktivně kódovány takovým způsobem, aby bylo možné je převést do podoby kvantitativních dat. Deduktivní kódování bude použito, protože výsledky budou srovnávány s jinou studií, a jejich závěry by měly být nezávislé. Důraz bude kladen zejména na interakci mezi popisnou a hodnotící částí výpovědi. Kódy budou zaznamenávány tak, aby byla patrná struktura výpovědi, tedy návaznosti mezi jednotlivými kódy. Výsledkem kódování, tak budou dvě vzájemně propojené vrstvy, obsahová a strukturní. V dalším postupu je mým záměrem využít možností shlukové analýzy s předpokladem, že budou nalezeny skupiny výpovědí obsahující podobnou tématiku a sktrukturu, tedy podobné vazby mezi kódy samotné výpovědi v rozsahu, který by pouze kvalitativní výzkum nedovoloval. Cílem je odhalit skupiny respondentů, které jsou si podobné ve více aspektech toho, jak o opilosti vypovídají.

Vzhledem k tomu, že si tento design výzkumu neklade za cíl poskytnout závěry kvantitativního charakteru, bude sběr dat probíhat metodou *snow ball* do teoretického nasycení dat. Míra saturace dat se ale plně projeví až při kvantitativní analýze. Předpokládaná velikost vzorku je zhruba v rozsahu 60 respondentů. Přechod ke kvantitativní analýze může být v menší míře proveden několikrát s případným doplněním dat v případě potřeby.

Mým cílem je srovnat závěry získané za pomoci tohoto *mixed method* designu, se závěry klasického kvalitativního výzkumu, a tak odhalit benefity a limity této metody. Evaluace výzkumného procesu, bude hlavním výstupem mé práce.

Předběžná struktura práce

1. Úvod – *mixed method* design
2. Téma výzkumu – opilost v životě mládeže
3. Metodologická část – popis zvolené metody
4. Praktická část – aplikace metody
5. Závěr – zhodnocení metody
6. Diskuse

Orientační seznam literatury

Brannen, J., 2016. *Mixing Methods: qualitative and quantitative research*, Oxon: Routledge.

Buchtík, M., 2016. *Mladí lidé a alkohol: závěrečná zpráva z výzkumu*, Sociologický ústav AV ČR.

Cortazzi, M., 2002. *Narrative Analysys*, Oxon: Routledge.

Creswell, J.W., c2014. *Research design: qualitative, quantitative, and mixed methods approaches* 4th ed., Thousand Oaks: SAGE Publications.

Everitt, B.S. et al., 2011. *Cluster Analysis* 5th edition., London: King's College.

Hájek, M., 2014. *Čtenář a stroj: vybrané metody sociálněvědní analýzy textů*, Praha: Sociologické nakladatelství (SLON).

Mareš, P., Rabušic, L. & Soukup, P., 2015. *Analýza sociálněvědních dat (nejen) v SPSS*, Brno: Masarykova univerzita.

Morse, J.M. & Niehaus, L., 2016. *Mixed Method Design: Principles and Procedures*, New York: Routledge.

Pearce, L.D., 2012. Mixed Methods Inquiry in Sociology. *American Behavioral Scientist*, 56(6), pp.829-848.

Strauss, A.L. & Corbin, J.M., 1999. *Základy kvalitativního výzkumu: postupy a techniky metody zakotvené teorie*, Brno: Sdružení Podané ruce.

Jméno konzultanta a jeho písemný souhlas se spoluprací

_____ _____

doc. Mgr. Martin Hájek, Ph.D. podpis studenta: Viktor Jurdič

Přílohy

Stop slova

a, aby, ačkoli, ač, aha, akorát, ahoj, ale, anebo, ano, aneb, asi, ani, aniž, aspoň, alespoň, apod, atp, atd, ať, až, během, bez, beztak, blízko, bohužel, brzy, být, býtli, buď, buďto, byť, čau, část, částečně, často, častý, či, člověk, cca, chtít, chuť', chuť, co, coby, copak, cokoli, cosi, což, čtrnáct, čtyři, daleko, dále, další, daný, děkovat, den, denně, deset, desítka, desetkrát, devatenáct, devět, díky, dnes, dle, do, dobrý, docela, dokonce, dole, doslova, druhý, dva, dvojka, dvacet, dvanáct, ehm, ergo, hodně, já, jakoby, jako, jaký, jakýsi, jak, jakýkoli, jakkoliv, jakož, jakože, i, jakmile, jaksi, jít, jeden, jedenáct, jednak, jednou, jet, jeho, jelikož, on, jen, jenž, jenže, jenom, ještě, jestli, jestliže, jinak, jiný, jo, k, kam, každopádně, kde, kdežto, kdesi, kdo, kdy, kdykoli, kdyby, kdysi, když, kolik, konkrétně, kromě, který, kvůli, kupodivu, leccos, leckdo, leč, leda, lze, mít, málo, můj, mezitím, mezi, mimo, mimochodem, moc, moci, možná, možný, muset, na, nad, nakonec, napodruhé, naposled, naproti, například, naopak, ne, něco, nebo, dělat, nějak, nějaký, několik, nejen, někde, někdo, někdy, některý, nejprve, stačit, vadit, než, nic, nijak, nikoli, nikterak, nula, obvykle, od, odtud, o, oba, opak, opět, osm, ovšem, osmnáct, pak, pakliže, pan, paní, patnáct, pět, po, podle, podobně, pokud, poněkud, poté, potažmo, pořád, pouze, potom, protože, proto, pozdě, prvý, před, předem, především, přece, přes, přesto, přestože, při, přičemž, přitom, příklad, právě, pro, proč, prosit, prostě, proti, první, rámec, rok, rovně, s, samozřejmě, samý, se, sice, sem, sedm, sedmnáct, šest, šestnáct, skoro, smět, snad, spíše, spolu, současně, svůj, sto, stý, ten, tady, tak, takto, také, takhle, taky, takový, takže, tam, tamhle, tamhleten, tamten, tento, ty, ted', tedy, téměř, tenhle, teprve, tisíc, tj, totiž, třeba, tři, třetí, třikrát, třináct, trošku, tvůj, určitě, u, úplně, už, v, věc, večer, vedle, velmi, velice, včera, vlastně, včetně, však, všechen, vůbec, vůči, vysoko, vždy, vždycky, vždyť, za, zejména, zač, zatím, zatímco, zase, zároveň, z, zas, zcela, zda, zjevně, zrovna, zřejmě, že, žel