Abstract

This thesis analyzes various text classification techniques in order to assess whether the knowledge of published news articles about selected companies can improve its' stock return volatility modelling and forecasting. We examine the content of the textual news releases and derive the news sentiment (polarity and strength) employing three different approaches: supervised machine learning Naive Bayes algorithm, lexicon-based as a representative of linguistic approach and hybrid Naive Bayes. In hybrid Naive Bayes we consider only the words contained in the specific lexicon rather than whole set of words from the article. For the lexicon-based approach we used independently two lexicons one with binary another with multiclass labels. The training set for the Naive Bayes was labeled by the author. When comparing the classifiers from the machine learning approach we can conclude that all of them performed similarly with a slight advantage of the hybrid Naive Bayes combined with multiclass lexicon. The resulting quantitative data in form of sentiment scores will be then incorporated into GARCH volatility modelling. The findings suggest that information contained in news feeds does bring an additional explanatory power to traditional GARCH model and is able to improve it's forecast. On the contrary, we could not provide enough evidence for favouring specific sentiment-derivation method. While the model employing hybrid Naive Bayes approach provided a bitter in-sample fit, the preferred model in the out-of-sample evaluation was the one employing multiclass lexicon. We also showed an asymmetric news effect, where both positive and negative news increase volatility with a latter having a more pronounced effect.

JEL Classification	C22, C52, C58, G14, G17, G41
Keywords	volatility, text, classifier, lexicon, sentiment, news

Author's e-mail	ksenia1105@gmail.com
Supervisor's e-mail	boril.sopov@gmail.com