Exploratory analysis of news sentiment using subgroup discovery

Anita Valmarska
Jožef Stefan Institute
Jamova cesta 39
and University of Ljubljana,
Faculty of Computer
and information science
Večna pot 113
Ljubljana, Slovenia
anita.valmarska@ijs.si

Luis Adrián Cabrera-Diego
and **Elvys Linhares Pontes**
L3i laboratory
University of La Rochelle
La Rochelle, France
luis.cabrera_diego,
elvys.linhares_pontes
{@univ-lr.fr}

Senja Pollak
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
senja.pollak@ijs.si

Abstract

In this study, we present an exploratory analysis of a Slovenian news corpus, in which we investigate the association between named entities and sentiment in the news. We propose a methodology that combines Named Entity Recognition and Subgroup Discovery - a descriptive rule learning technique for identifying groups of examples that share the same class label (sentiment) and pattern (features - Named Entities). The approach is used to induce the positive and negative sentiment class rules that reveal interesting patterns related to different Slovenian and international politicians, organizations, and locations.

1 Introduction

Traditionally, sentiment analysis refers to the use of natural language processing to systematically identify, extract, quantify, and study affective states and subjective information. Most frequently, it is used as a predictive technique used to model social media (Beigi et al., 2016), more specifically to predict or summarize opinions, attitudes and emotions in tweets, comments, online reviews etc., where the main focus is on predicting attitudes expressed towards a specific entity (Mejova, 2009). Another line of research applies sentiment analysis on news text, where the focus has shifted from analyzing sentiment towards a specific target to analyzing the intrinsic mood of the text itself (Pelicon et al., 2020). Authors aimed to model feelings (positive, negative, or neutral) that readers feel while reading a certain piece of news (Bučar et al., 2018; Liu, 2012; Pelicon et al., 2020), also in relation to news covering Covid-19 (Aslam et al., 2020), modelled news sentiment in relation to stock market and economic conditions (Van de Kauter et al., 2015; Bowden et al., 2019; Rambaccussing and Kwiatkowski, 2020). Sentiment analysis has been also used in fake news identification (Bhutani et al., 2019) and in media bias analysis (El Ali et al., 2018).

In the current trend of natural language processing research (Rogers and Augenstein, 2020), the main focus is on improving the predictive performance over state-of-the art especially using deep learning-based methods. The drawback of these models is in their very limited interpretability. In contrast, several data and text mining techniques have been developed to improve domain understanding and support exploratory analysis of data, with focus on explainable models, which is crucial e.g. in medical applications, but also interesting for interdisciplinary research in the field of digital humanities and digital social sciences. Our research falls under this line of research.

The aim of our study is to gain better understanding into news sentiment by analysis of named entities in a manually annotated corpus of Slovenian news articles (Bučar, 2017). More specifically, our aim is to identify groups of topics with negative or positive sentiment in Slovenian news, where topics are identified by named entities and their interaction forms the context of the reported stories. We propose the employment of subgroup discovery — a descriptive rule learning technique for identification of groups of examples sharing the same class label (sentiment) and same pattern (features). The task of subgroup discovery is the combination of predictive and descriptive rule induction. The result of subgroup discovery is to provide understandable descriptions of subgroups of individuals which share a common target property of interest. Subgroup discovery methods have traditionally be successfully applied to in different medical applications (e.g. detecting of groups of patients at risk for atherosclerotic cardiovascular disease (Gamberger and Lavrač, 2002), supporting factors for brain ischemia (Gamberger and Lavrač, 2007), and psychiatric emergency (Carmona et al.,

Proceedings of the 8th BSNLP Workshop on Balto-Slavic Natural Language Processing, pages 66–72
April 20, 2021.

2011), but only rarely applied to model textual data. The closest to our study is the work by (Vavpetič et al., 2013) using the subgroup discovery system Hedwig for analyzing news articles about Portugal focusing on interesting vocabulary patterns that reflect credit default swap. The authors focused on financial entities, geographical entities and a specialized vocabulary of the European sovereign debt crisis.

The main contributions of this paper are two fold. First, we propose a novel approach using named entity recognition and linking in a subgroup discovery setting. Next, we apply the method on the Slovenian news dataset, getting new insights into Slovenian news reporting in terms of news sentiment and showcase the potential of our approach for digital social science research.

The paper is structured as follows: in Section 2, we present the data used in the experimental work. Section 3 presents a short outline of the employed methodology. In Section 4 and Section 5, we present our results and offer our conclusions and ideas for further work.

2 Data

In our experiments, we used the manually sentiment annotated Slovenian news corpus SentiNews 1.0 (Bučar et al., 2018; Bučar, 2017)[1]. The corpus consists of Slovene web-crawled news containing more than 250,000 documents with political, business, economic and financial content from five Slovene media resources on the web. The data covers the period between 1 September 2007 to 31 December 2013. Data used in the experiments is a manually sentiment annotated stratified random sample of 10,427 documents from news portals 24ur, Dnevnik, Finance, Rtvslo, and Žurnal24. Data was independently annotated by 2-6 annotators, using the five-level Lickert scale (1 – very negative, 2 – negative, 3 – neutral, 4 – positive, and 5 – very positive) on three levels of granularity, i.e. on document, paragraph, and sentence level. The sentiment of an instance is defined as the average of the sentiment scores given by the different annotators, where an instance labeled as negative has received an average score less than or equal to 2.4 and an instance labeled as positive has received an average score to 3.6. Instances with an average score in-between were labeled as neutral. The analysis of the agreement between annotators is available in (Bučar et al., 2018). The value of annotators agreement on document level as measured by the Cronbach's alpha is 0.903.

In this paper we are interested only in documents with either positive or negative sentiment, which corresponds to 1665 positive and 3337 negative articles, respectively. Note that the dataset is thus imbalanced towards the negative class (which is also matching the observations of media researchers that attention to negative news is disproportionate (e.g. (Van der Meer et al., 2019; Soroka et al., 2019)).

3 Methodology

The methodology for named entity-based sentiment subgroup discovery consists of three steps.

3.1 Named entity recognition and linking

For each document from the corpora, we perform named entity recognition (NER) and named entity linking (NEL) using the approaches described in Boros et al. (2020)[2] and Linhares Pontes et al. (2020)[3], respectively. Specifically, for the NER system we fine-tuned CroSloEngual BERT (Ulčar and Robnik-Šikonja, 2020) with two staked Transformer blocks on the top. For the NEL system we used the architecture founded on the Multilingual End-to-End Entity Linking with match correction and candidate filtering. Both systems, NER and NEL, were trained using the Slovene WikiANN dataset (Pan et al., 2017). The dataset was split in three partitions, train, development and test. The evaluation on the test partition, showed that the NER system has a micro F-score of 0.954, while the NEL system has an F-score of 0.705.

In SentiNews 1.0, we identified 914 person names, 699 organizations and 476 locations with assigned NEL identifiers. We used the NEL codes to extract the nominative case of the named entities from the the Slovenian Wikipedia.

3.2 Data transformation

As the state-of-the-art algorithms for subgroup discovery work on structured data, the second step of the methodology is to transform the discovered (and linked) named entities from step 1 into a tabular form suitable for subgroup discovery. The resulting tables were constructed by representing

[1]Data is available on https://www.clarin.si/repository/xmlui/handle/11356/1110

[2]Code available at: https://github.com/EMBEDDIA/stacked-ner

[3]Code available at: https://github.com/EMBEDDIA/multilingual_entity_linking

each non-neutral sentiment document from the corpora as a row in a table. The documents are described by the values of the identified entities, *yes* if the respective entity was identified in the document and *no* if the entity was not present in the document. The document's sentiment represent the class label.

The result of data transformation is a table with 2645 rows (i.e. documents with positive or negative class label, and identified linked named entity) and 2089 columns (i.e. named entities corresponding to person, organisation and location names) as attributes. In our order to reduce the number of columns and improve the chances of the subgroup discovery algorithm of discovering good subgroups, we chose to proceed only with top n (with n=20) most frequent entities from each entity group: person, organization, and location.[4] As we are only interested into entities mentioned in the articles, we removed all documents without any identified entity, thus resulting in a table with 1703 rows and 60 columns i.e. entities, with the positive and negative sentiment class distribution (560, 1143), which is similar to the dataset distribution before preprocessing).

Table 1 presents the 20 most frequent persons, organizations, and locations identified in the corpora. Entities are ordered according to their frequency, the most popular entities are written on top.

Persons	Organizations	Locations
Borut Pahor	Ljubljanska borza	Združeno kraljestvo
Janez Janša	Evropska komisija	Luka Koper
Dow Jones	Evropska unija	New York
Danilo Türk	Telekom Slovenije	Nova Gorica
Igor Bavčar	Košarkarski klub Zlatorog	Murska Sobota
Karl Erjavec	Radiotelvizija Slovenija	Ljubljanska borza
Gregor Virant	Newyorška borza	Slovenj Gradec
Alenka Bratušek	Luka Koper	Mestna občina Ljubljana MOL
Katarina Kresal	Adria Airways	Slovenija
Nogometni klub Koper	Dow Jones	Črna gora
Mitja Gaspari	Nova Ljubljanska banka NLB	Novo mesto
Angela Merkel	Wall Street	Škofja Loka
Goldman Sachs	Svetovna banka	Kranjska Gora
Gregor Golobič	Pop TV	Bosna in Hencegovina
Nicolas Sarkozy	General Motors	Bela krajina
Ivo Sanader	Lehman Brothers	Južna Koreja
Dimitrij Rupel	Bank of America	Hrvaška
Andrej Bajuk	Republika Slovenija	Mestna občina Maribor MOM
Janeza Drnovška	Standard & Pool	Nova Zelandija
Barack Obama	Deutsche Bank	Grad Brdo pri Kranju

Table 1: Twenty most frequent persons, organizations, and locations in SentiNews1.0.

3.3 Subgroup discovery

The third step of the methodology is the identification of subgroups via subgroup discovery. Subgroup discovery is a descriptive induction technique that learns descriptive rules from labeled data. The task of subgroup discovery is to find interesting subgroups in the population, i.e. subgroups that have a significantly different class distribution than the entire population (Klösgen, 1996; Wrobel, 1997). The result of subgroup discovery is a set of individual rules, where the rule consequence is a class label. An important characteristic of subgroup discovery is that its task is a combination of predictive and descriptive rule induction. It provides understandable descriptions of subgroups of individuals which share a common target property of interest.

We performed subgroup discovery using the DoubleBeam-SD algorithm (Valmarska et al., 2017)[5]. The DoubleBeam-SD subgroup discovery algorithm combines separate refinement and selection heuristics with the beam search. Inverted m-estimate is used as the heuristics in the rule refinement phase while m-estimate is the rule selection heuristics. The width of the beam was set to 100. The algorithm was set to extract up to 10 best rules for each class.

4 Results

Table 2 presents the rules describing groups of documents with positive or negative sentiment. The prefixes *PER_*, *ORG_*, and *LOC_* denote a person, an organization or a location respectively. For example, the interpretation of the first rule, `LOC_Nova_Gorica = no AND PER_Igor_Bavčar = yes AND LOC_Grad_Brdo_pri_Kranju = no` $\rightarrow$ `negative` is as follows: articles that talk about the person Igor Bavčar and do not talk about the locations *Nova_Gorica* and *Grad_Brdo_pri_Kranju* (corresponding to Brdo pri Kranju, the location of

[4]The number of most frequent named entities (n) from each category was set arbitrarily.

[5]The code of the DoubleBeam-SD algorithm is available on https://github.com/bib3rce/RL_SD.

Slovenian government's main venue for diplomatic meetings) are articles with *negative* sentiment. This rule covers 28 documents, out of which 26 are actually labeled with *negative* sentiment, while 2 are annotated with positive sentiment.

Rules describing articles with *negative* sentiment concern articles written about Igor Bavčar, a Slovenian politician and manager. As politician, he played an important role during the Slovenian independence period (he acted as minister of internal affairs in the beginning of the 90s and later as Slovenian minister for European affairs) (Pluut-Pregelj et al., 2018). In 2002, he withdraw from politics to engage in the private sector and was connected to financial affairs, which explains the negative sentiment association in our corpus. Bavčar became the chairman of the Istrabenz holding company in 2002. On 31 March 2009 Bavčar resigned as the President of Istrabenz due to poor business performance as the company was forced into financial restructuring. Bavčar was indicted for disputed trading in Istrabenz shares in 2017, when he attempted to take over the company. He was arrested in financial fraud investigation and released after 10 hours.[6]

The discovered subgroups of articles with *positive* sentiment contain articles talking either about Slovenia's former president, Danilo Türk, or the city of New York, *New Yorku*. Danilo Türk is a Slovenian diplomat, professor of international law, human rights expert, and political figure who served as President of Slovenia from 2007 to 2012. Note that the rules of discovered subgroups have a low precision (as the distribution of true positive and false positive is 53% vs. 47%). However one should take into account that the dataset is very imbalanced — with only 33% of articles annotated with a positive sentiment class label. These rules also show the main objective of subgroup discovery — discovering and describing groups of examples with significantly different class distribution to the the population class distribution rather than prediction.

It is interesting that the positive sentiment for New York is connected to the absence of other named entities, closely related to New York including the former president of the United States, Barack Obama. Another entity whose omission in news articles strengthens the positive sentiment associated with New York is the investment bank Lehman Brothers. On September 15, 2008, Lehman Brothers Holdings, Inc. sought Chapter 11 protection, initiating the largest bankruptcy proceeding in United States (U.S.) history. It declared $639 billion in assets and $613 billion in debts. At the time, Lehman was the fourth-largest U.S. investment bank. Despite being thought "too big to fail", the federal government did not employ extraordinary measures to save Lehman, such as the enabling financing it had facilitated for J.P. Morgan Chase's purchase of a failing Bear Stearns just six months earlier. Lehman's demise was a seminal event in the financial crisis that began in the U.S. subprime mortgage industry in 2007, spread to the credit markets, and then burned through the world's financial markets. The crisis resulted in significant and wide losses to the economy. (Wiggins et al., 2014)

Other rules of positive sentiment with named entity New York, also include the absence of Standard & Poor's, an American credit rating agency and a division of S&P Global (*ORG_Standard_&_Pools*) that publishes financial research and analysis on stocks, bonds, and commodities, and the Bank of America (*ORG_Bank_of_America*).

The positive sentiment of Danilo Türk is related to the omission of the above mentioned politician involved in financial affairs, Igor Bavčar. The quality of the rules presented in Table 2 is presented in Table 3 in the appendix.

5 Conclusions

In this paper, we presented a sentiment subgroup discovery approach through the lenses of named entities. Experiments were performed on a corpora of Slovene news data, manually labeled with positive and negative sentiment. Our approach utilizes the fact that the sentiment in news articles is tied-up to their discussed named entities. We identified groups of news articles with negative and positive sentiment.

The identified groups of articles were described by rules containing only one positive condition, confirming the presence of a named entity in an article. Many of the identified subgroups for a chosen sentiment focus on the same entity. One of the key person entities associated with negative sentiment is Igor Bavčar - a polarizing Slovenian politician and manager, whose presence in the news from 2009 is accompanied by negative sentiment due to

[6] https://en.wikipedia.org/wiki/Igor_Bav%C4%8Dar, Last accessed: 8 February 2021.

R.nr.	Rule	Sentiment	TP	FP
N1	LOC_Nova_Gorica = no AND PER_Igor_Bavčar = yes AND LOC_Grad_Brdo_pri_Kranju = no	→ negative	26	2
N2	PER_Igor_Bavčar = yes AND ORG_Telekom_Slovenije = no AND LOC_Grad_Brdo_pri_Kranju = no AND PER_Dimitrij_Rupel = no	→ negative	26	2
N3	LOC_Hrvaška = no AND PER_Igor_Bavčar = yes AND LOC_Grad_Brdo_pri_Kranju = no AND PER_Dimitrij_Rupel = no	→ negative	26	2
N4	PER_Ivo_Sanader = no AND PER_Igor_Bavčar = yes AND LOC_Grad_Brdo_pri_Kranju = no	→ negative	26	2
N5	PER_Igor_Bavčar = yes AND LOC_Grad_Brdo_pri_Kranju = no AND PER_Mitja_Gaspari = no AND PER_Dimitrij_Rupel = no	→ negative	26	2
N6	PER_Igor_Bavčar = yes AND ORG_Bank_of_America = no	→ negative	26	2
N7	PER_Igor_Bavčar = yes AND LOC_Grad_Brdo_pri_Kranju = no AND LOC_New_York = no	→ negative	26	2
N8	LOC_Nova_Zelandija = no AND PER_Igor_Bavčar = yes AND LOC_Grad_Brdo_pri_Kranju = no AND PER_Dimitrij_Rupel = no	→ negative	26	2
N9	LOC_Slovenija = no AND PER_Igor_Bavčar = yes AND LOC_Grad_Brdo_pri_Kranju = no	→ negative	26	2
N10	PER_Ivo_Sanader = no AND PER_Igor_Bavčar = yes	→ negative	26	2
P1	ORG_Nova_Ljubljanska_banka_NLB = no AND PER_Angela_Merkel = no AND LOC_New_York = yes	→ positive	57	51
P2	PER_Barack_Obama = no AND ORG_Evropska_komisija = no AND LOC_New_York = yes	→ positive	57	51
P3	ORG_Nova_Ljubljanska_banka_NLB = no AND PER_Igor_Bavčar = no AND PER_Danilo_Türk = yes	→ positive	24	18
P4	ORG_Lehman_Brothers = no AND ORG_Evropska_komisija = no AND LOC_New_York = yes	→ positive	57	51
P5	PER_Igor_Bavčar = no AND PER_Angela_Merkel = no AND LOC_New_York = yes	→ positive	57	51
P6	ORG_Standard_&_Pool = no AND ORG_Evropska_komisija = no AND LOC_New_York = yes	→ positive	57	51
P7	ORG_Nova_Ljubljanska_banka_NLB = no AND PER_Igor_Bavčar = no AND ORG_Košarkarski_klub_Zlatorog = no AND PER_Danilo_Türk = yes	→ positive	24	18
P8	ORG_Košarkarski_klub_Zlatorog = no AND PER_Angela_Merkel = no AND LOC_New_York = yes	→ positive	57	51
P9	ORG_Bank_of_America = no AND ORG_Evropska_komisija = no AND LOC_New_York = yes	→ positive	57	51
P10	ORG_Nova_Ljubljanska_banka_NLB = no AND PER_Igor_Bavčar = no AND PER_Danilo_Türk = yes AND ORG_Združeno_kraljestvo = no	→ positive	57	51

Table 2: List of rules describing subgroups of articles with *negative* and *positive* sentiment.

financial affairs. We are also able to identify a generally positive sentiment towards another Slovenian politician and a location.

In future work we will adapt the used algorithm for subgroup discovery to identify subgroups that are not overlapping, thus potentially involving multiple entities.

Acknowledgments

This work was supported by the Slovenian Research Agency (ARRS) grants for the programmes, Knowledge technologies (P2-0103) and Artificial intelligence and inteligent systems (P2-0209), the project Computer-assisted multilingual news discourse analysis with contextual embeddings (CANDAS, J6-2581), as well as the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

Faheem Aslam, Tahir Mumtaz Awan, Jabir Hussain Syed, Aisha Kashif, and Mahwish Parveen. 2020. Sentiments and emotions evoked by news headlines of coronavirus disease (covid-19) outbreak. *Humanities and Social Sciences Communications*, 7(1):1–9.

Ghazaleh Beigi, Xia Hu, Ross Maciejewski, and Huan Liu. 2016. An overview of sentiment analysis in social media and its applications in disaster relief. *Sentiment analysis and ontology engineering*, pages 313–340.

Bhavika Bhutani, Neha Rastogi, Priyanshu Sehgal, and Archana Purwar. 2019. Fake news detection using sentiment analysis. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pages 1–5. IEEE.

Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Alleviating Digitization Errors in Named Entity Recognition for Historical Documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL)*, pages 431–441, Online. Association for Computational Linguistics.

James Bowden, Andrzej Kwiatkowski, and Dooruj Rambaccussing. 2019. Economy through a lens: Distortions of policy coverage in uk national newspapers. *Journal of Comparative Economics*, 47(4):881–906.

Jože Bučar. 2017. Manually sentiment annotated slovenian news corpus SentiNews 1.0. Slovenian language resource repository CLARIN.SI.

Jože Bučar, Martin Žnidaršič, and Janez Povh. 2018. Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, 52(3):895–919.

Cristóbal J Carmona, Pedro González, María José del Jesus, Mercedes Navío-Acosta, and Luis Jiménez-Trevino. 2011. Evolutionary fuzzy rule extraction

for subgroup discovery in a psychiatric emergency department. *Soft Computing*, 15(12):2435–2448.

Abdallah El Ali, Tim C Stratmann, Souneil Park, Johannes Schöning, Wilko Heuten, and Susanne CJ Boll. 2018. Measuring, understanding, and classifying news media sympathy on twitter after crisis events. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Dragan Gamberger and Nada Lavrač. 2007. Supporting factors in descriptive analysis of brain ischaemia. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 155–159. Springer.

Dragan Gamberger and Nada Lavrač. 2002. Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, 17:501–527.

Marjan Van de Kauter, Diane Breesch, and Véronique Hoste. 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with applications*, 42(11):4999–5010.

Willi Klösgen. 1996. Explora: A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*, pages 249–271. AAAI/MIT Press.

Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupco Todorovski. 2004. Subgroup discovery with cn2-sd. *J. Mach. Learn. Res.*, 5(2):153–188.

Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Emanuela Boros, Ahmed Hamdi, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2020. Entity Linking for Historical Documents: Challenges and Solutions. In *Proceedings of the 22nd International Conference on Asia-Pacific Digital Libraries (ICADL 2020)*, pages 215–231, Kyoto, Japan. Springer International Publishing.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Toni GLA Van der Meer, Anne C Kroon, Piet Verhoeven, and Jeroen Jonkman. 2019. Mediatization and the disproportionate attention to negative news: The case of airplane crashes. *Journalism Studies*, 20(6):783–803.

Yelena Mejova. 2009. Sentiment analysis: An overview. *University of Iowa, Computer Science Department*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Andraž Pelicon, Marko Pranjić, Dragana Miljković, Blaž Škrlj, and Senja Pollak. 2020. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10(17):5993.

Leopoldina Plut-Pregelj, Gregor Kranjc, Žarko Lazarević, and Carole Rogel. 2018. *Historical Dictionary of Slovenia (Historical Dictionaries of Europe)*. Rowman & Littlefield Publishers.

Dooruj Rambaccussing and Andrzej Kwiatkowski. 2020. Forecasting with news sentiment: Evidence with uk newspapers. *International Journal of Forecasting*, 36(4):1501–1516.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in nlp? *arXiv preprint arXiv:2010.03863*.

Stuart Soroka, Patrick Fournier, and Lilach Nir. 2019. Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, 116(38):18888–18892.

Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT. In *Text, Speech, and Dialogue*, pages 104–111, Cham. Springer International Publishing.

Anita Valmarska, Nada Lavrač, Johannes Fürnkranz, and Marko Robnik-Šikonja. 2017. Refinement and selection heuristics in subgroup discovery and classification rule learning. *Expert Systems with Applications*, 81:147–162.

Anže Vavpetič, Petra Kralj Novak, Miha Grčar, Igor Mozetič, and Nada Lavrač. 2013. Semantic data mining of financial news articles. In *International Conference on Discovery Science*, pages 294–307. Springer.

Rosalind Wiggins, Thomas Piontek, and Andrew Metrick. 2014. The Lehman Brothers bankruptcy a: Overview. *Yale program on financial stability case study*.

Stefan Wrobel. 1997. An algorithm for multi-relational discovery of subgroups. In *Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery, PKDD 1997*, pages 78–87.

A Appendix. Quality of subgroup rules

Table 3 presents the quality measures of rules describing subgroups presented in Table 2. Note that the purpose of these experiments is exploratory analysis, not a prediction classification task and that the results are reported on the entire dataset, without any train-test data split. In addition to the traditional measures of precision, accuracy and recall, we also present values of *weighted relative accuracy* (WRACC).

R.nr	Precision	Accuracy	Recall	WRACC
N1—N10	0.929	0.685	0.023	0.004
P1—P2	0.528	0.332	0.102	0.013
P3	0.571	0.332	0.043	0.006
P4—P6	0.528	0.332	0.102	0.013
P7	0.571	0.332	0.043	0.006
P8—P10	0.528	0.332	0.102	0.013

Table 3: Quality of the rules describing subgroup presented in Table 2.

Lavrač et al. (2004) argue that WRACC, also referred to as *unusualness*, is the most important measure of subgroup discovery rules. WRACC for a chosen rule is defined as follows

$$WRACC = \frac{TP + FP}{P + N}\{\frac{TP}{TP + FP} - \frac{P}{P + N}\}, \quad (1)$$

where TP is the number of *true positive* examples covered by the rule, FP is the number of *false positive* examples covered by the rule, while P and N are the number of positive and negative examples, respectively, in the population. WRACC reflects both the rule significance and rule coverage, as subgroup discovery is interested in rules with significantly different class distribution than the prior class distribution that covers many instances.