



Title	Sentiment Analysis of Polish Online News Covering Controversial Topics : Comparison Between Lexicon and Statistical Approaches
Author(s)	Szwoch, Joanna; Staszko, Mateusz; Rzepka, Rafal et al.
Citation	Human Language Technologies as a Challenge for Computer Science and Linguistics, 2023, 277-281
Issue Date	2024-02-06
Doc URL	https://hdl.handle.net/2115/95945
Rights(URL)	https://creativecommons.org/licenses/by-nc-nd/4.0/
Type	conference paper
File Information	Proc. Hum. Lang. Technol. Chall. Comput. Sci. Linguist. 2023_277-281.pdf



Sentiment Analysis of Polish Online News Covering Controversial Topics – – Comparison Between Lexicon and Statistical Approaches

Joanna Szwoch¹, Mateusz Staszko², Rafal Rzepka³, Kenji Araki³

¹Graduate School of Information Science and Technology, Hokkaido University,
joannaeleonora.szwoch.u0@elms.hokudai.ac.jp

²Mateusz Staszko Software Development, mateuszstasz@gmail.com

³Faculty of Information Science and Technology, Hokkaido University, {rzepka, araki}@ist.hokudai.ac.jp

Abstract

In this paper we present results of a sentiment analysis performed on a part of the Polish Online News Corpus which consists of articles covering controversial topics in Poland. We perform both lexicon-based and statistical experiments and compare the obtained results for this task. The main goal of our comparative work is to discover whether supposedly objective online news contain any emotive content and how different the results are, depending on the approach. Our study found no meaningful differences in emotional load at the word level, but the multilingual language model trained with the Natural Language Inference objective has partially confirmed its existence. We believe this work can be used for further research on media polarization in Polish language.

Keywords: Polish language, online news, sentiment analysis

1. Introduction

The Internet is a place full of content generated by people. To a large extent, people's views and decisions that follow them are shaped by the content they encounter online (Burbach et al., 2020). Moreover, in present days, traditional media such as newspapers, magazines, radio or even TV are slowly fading away, in favour of news outlets available on the Internet. News providers, according to the highest journalistic standards, are supposed to deliver pure information without any bias. However, there are examples where this condition is not met (Elejalde et al., 2018). In this work we want to investigate if two popular online news providers in Poland are following the aforementioned standards.

News outlets which are considered biased, tend to present information in favour of a certain side, whether it is a political party or a social movement. Their promotion of a specific perspective can lead to a spread of confusing or inaccurate information, causing mistrust and divisions within a society (Aggarwal et al., 2020).

Sentiment analysis and emotion detection are ways to determine the emotions hidden in text and to gain insights on how people feel about certain topics (Nandwani and Verma, 2021). It can help to detect fake and biased news as well as misinformation. In general, the problem of including opinions and emotions in contemporary media spotlights the importance of critical thinking and literacy in current digital age (Machete and Turpin, 2020).

The goal of sentiment analysis is to identify subjective information within a text, such as opinions or emotions. There are two approaches to this task, namely a rule-based and a statistical-based one. The former one relies on manually-crafted rules, including dictionaries and corpora to classify text into categories. Three classes, that is positive, negative and neutral are examples of such categorization, but they can also consist of various emotions. On the other hand, the latter approach can be divided into three methods – supervised, unsupervised and self-supervised. Both approaches have strong points and limitations depending on the type of input data and the use case.

Polish language is rather neglected when it comes to the number of studies on sentiment analysis. In comparison with other languages, there is a relatively small number of available resources. It is an obstacle for researchers wishing to delve into this topic. Despite this challenge, in this paper we experiment with two different approaches to opinion mining in Polish language. We also discuss the pros and cons of the utilized techniques and potential future directions in this field.

The rest of this paper is structured as follows. In Section 2 we mention prior works on sentiment analysis, including different approaches to the task depending on the length of text and language. Section 3 describes utilized datasets and Section 4 outlines methods applied in our experiments. In Section 5 we analyze obtained results and in Section 6 we present points open for discussion. Section 7 focuses on conclusions which can be drawn from the conducted experiments and our future work plans.

2. Related works

Opinion mining became a fairly popular branch of natural language studies, which resulted in a bountiful of works covering the topic (Mäntylä et al., 2016; Cui et al., 2023). Majority of studies relate to short inputs such as tweets or reviews. In our work we will focus on examining emotions in news, which is a fairly uncommon approach.

One of the examples of lexicon-based approaches is a study which performed sentiment analysis on English BBC news (Shirsat et al., 2017). Bing sentiment dictionary allowed

classification of positive, negative or neutral content. Such approach showed that in different news categories information tends to be polarized rather than impartial. Another popular lexicon-based technique includes VADER¹. It is prepared for English vocabulary, but a common practice is to translate this source to any desired language, as in the study of sentiment analysis of health-related online news in Brazilian Portuguese (Mello et al., 2022). Other examples of dictionaries include WordStat Sentiment Dictionary, SentiWordNet, AFINN, Loughran & McDonald Financial Sentiment Dictionary (L&M), Lexicoder Sentiment Dictionary (Tomanek, 2014). Out of these, only SentiWordNet includes a dedicated Polish version – plWordNet (Kocoń et al., 2017).

Sentiment analysis conducted on online financial texts indicates that RoBERTa-based model outperformed results obtained by SVM, Logistic Regression or Naive Bayes (Zhao et al., 2020). Moreover, opinion mining of Italian reviews indicated that machine and deep learning approaches such as BERT perform better in general, however, lexicon-based methods are recommended for small datasets with limited resources (Catelli et al., 2022).

Polish language studies include sentiment analysis of conversation transcripts with a social robot (Probiez and Galuszka, 2022). For this type of long text, lexicon-based approach was applied. Two different dictionaries – English NRC (Mohammad and Turney, 2013) and Polish plWordNet were used for this task. Both of them allow to associate words with 8 different types of emotions. However, due to the limited number of analyzed conversations, the variation between results was not observed.

On the other hand, multilingual BERT model yielded good results in sentiment analysis of news in another Slavic language, namely Slovenian (Pelicon et al., 2020). Moreover, this zero-shot learning approach was also used on Croatian language for the same task and it outperformed the results of other classifying models.

However, no similar experiments were conducted on Polish online news.

3. Datasets

In this section we describe the datasets used for sentiment analysis of news covering controversial topics in Polish language.

3.1. NAWL – The Nencki Affective Word List

The Nencki Affective Word List (NAWL) is a database consisting of 2,902 Polish words such as nouns, verbs, and adjectives (Riegel et al., 2015). All of them are rated for emotional valence, arousal, and imageability. Furthermore, measures of frequency, grammatical class, and number of letters are also included in this set. NAWL is a Polish version of the Berlin Affective Word List - Reloaded (Võ et al., 2009), which was created for the same purpose, but in German language. This dictionary describes each word with a point value in terms of 5 categories: anger, disgust, fear, happiness and sadness. Additionally, neutral and unknown

classes are included for words that do not have a dominant category. The higher the value, the stronger the word expresses the particular emotion.

NAWL is used in our non-statistical approach to sentiment analysis of words which appeared in examined online news. We chose this lexicon due to its accessibility and freeware licence.

3.2. Polish online news articles covering controversial topics

In this paper we used the Polish Online News Corpus to retrieve articles covering contemporary controversial topics in Poland (Szwoch et al., 2022). The original dataset consists of more than 200,000 online news articles in Polish language between years 2019 and 2021 from two major TV news broadcasters – TVP Info and TVN 24. As our work primarily focuses on performing sentiment analysis of articles related to subjects considered controversial, we extracted only relevant examples from the corpus. The topics were chosen arbitrarily and are related to contemporary topics which appear in Polish media. However, recent study on the political leaning discovery was an inspiration to consider topics such as EU and abortion (Baran et al., 2022).

The following subjects were selected as controversial ones in Polish society:

- Abortion: in Poland, it is only legal when pregnant woman's life is in danger or in case the pregnancy is a result of rape or incest. In 2021, it became illegal in case the fetus is severely malformed².
- Church: the Roman Catholic Church is the dominant religion in Poland.
- Constitution: recent constitution changes resulted in financial penalties from the European Union (EU)³.
- Presidential elections: last ones took place in 2020⁴.
- EU: Poland is a part of the EU. Citizens' attitude towards Poland's membership in the EU varies in time⁵.
- Independence March: it is an annual event that takes place in Warsaw on November 11th to commemorate regaining independence by Poland⁶.
- LGBT parade: the Equality Parade is an annual event that takes place in several cities in Poland to celebrate the LGBT community and promote tolerance and equality⁷.
- Refugees: the Polish-Belarusian border has become a site of increased activity in recent years, as refugees and migrants try to enter the EU via Poland⁸.

To select news articles on controversial topics, we filtered them using keywords that are listed in Table 1. We attempted to choose a non-biased wording. The keyword

²Dziennik Ustaw (Journal of Laws) Dz.U.2022.1575

³Order ECLI:EU:C:2021:878, Court of Justice of EU

⁴<https://wybory.gov.pl/index>

⁵https://www.cbos.pl/PL/trendy/trendy.php?trend_parametr=stosunek_do_integracji UE

⁶<https://marszniepodleglosci.pl/>

⁷<http://www.paradarownosci.eu/>

⁸<https://www.strazgraniczna.pl/pl/aktualnosci/9689,Nielegalne-przekroczenia-graniczy-z-Bialorusia-w-2021-r.html>

¹<https://github.com/rafjaa/LeIA>

“strike” for abortion was chosen due to its connection to the Polish Women’s Strike movement. To avoid repetitions of articles between categories, keywords from one subject were banned in others. Moreover, we limited website category to “Poland” to focus on matters within the country.

Topic	Keywords
Abortion	abortion, strike
Church	church, catholic, priest
Constitution	constitution, court, law, judge, tribunal
Presidential elections	elections
EU	union, european, cjeu
Independence March	independence, march, day
LGBT parade	lgbt, parade
Refugees	refugee, border, migrant

Table 1: Keywords used for selecting articles concerning each controversial topic

Table 2 shows the distribution of articles between topics. Analyses showed that the number of words in articles is resembling normal distribution. We decided to remove the outliers that were outside the three standard deviation distances from the mean length.

Topic	Number of articles	
	TVN	TVP
Abortion	62	143
Church	166	47
Constitution	147	92
Elections	234	205
EU	260	155
Independence March	46	86
LGBT parade	29	34
Refugees	124	147

Table 2: Number of articles in each topic per news outlet

4. Methods

This Section describes all methods of analyses performed in this work.

4.1. Lexicon-based classification with NAWL dictionary on articles covering controversial topics

In our first experiment sentiment analysis is conducted by matching words from selected articles, lemmatized with *SpacyPL*⁹ library, with pre-defined NAWL sentiment dictionary consisting of 6 classes. All words within an article are a subject of the analysis. The sentiment scores are aggregated to determine the sentiment of each topic per news provider. Two-tailed t-test is performed to compare the results between news providers depending on the topic and check if the differences are statistically significant.

4.2. Zero-shot classification on NAWL dictionary

In the second step we perform zero-shot classification of NAWL dictionary words with XLM-RoBERTa Large

⁹<https://spacy.io/models/pl>

model¹⁰, fine-tuned on XNLI multilingual dataset¹¹. We decided to use zero-shot approach as it yielded promising results for Slavic languages.

4.3. Zero-shot classification on articles covering controversial topics

Last experiment is a zero-shot classification on full article texts within each controversial topic. Six labels were provided in English, the same as in NAWL dictionary. Articles have been truncated to BERT-based models’ maximum capacity of 512 tokens. Obtained results were used to perform two-tailed t-test for any statistically significant differences between groups.

5. Results

In this section we present obtained results of performed experiments.

5.1. Lexicon-based classification with NAWL dictionary on articles covering controversial topics

Results of lexicon-based sentiment analysis on the article content are shown in the Table 4.

T-test carried out on all emotions showed that at 5% level of significance, we do not reject the null hypothesis on no difference between means for any categories, meaning that the differences between groups proved to be insignificant.

5.2. Zero-shot classification on NAWL dictionary

Results of RoBERTa classification on NAWL dictionary is shown in Table 3.

Emotion	NAWL	RoBERTa-XNLI
Anger	13.26	7.58
Disgust	6.50	39.51
Fear	22.06	2.841
Happiness	19.89	4.33
Neutral	29.63	42.35
Sadness	8.66	3.38

Table 3: Percentage of words in each emotion category, based on NAWL dictionary and RoBERTa-XNLI classification

Percentages showed in the Table 3 indicate that RoBERTa based model differently classifies most of the groups. It is worth mentioning that there are more words classified by RoBERTa as “Neutral” than in the original dictionary. Moreover, there is also a noticeable difference in a number of words associated with emotions “Disgust”, “Fear” and “Happiness”. The first one is almost 8 times bigger than in the original dictionary. The other two are respectively almost eleven and five times smaller than the lexicon.

5.3. Zero-shot classification on articles covering controversial topics

Results of lexicon-based sentiment analysis on article content are shown in the Table 5.

¹⁰<https://huggingface.co/xlm-roberta-large>

¹¹<https://huggingface.co/joeddav/xlm-roberta-large-xnli>

Emotion	Abortion		Church		Constitution		Elections		EU		Independence March		LGBT		Refugees	
	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP
Anger	13.88	10.40	5.81	5.84	4.68	4.76	4.57	5.12	4.38	4.30	5.65	5.66	6.22	5.87	5.34	5.32
Disgust	1.37	0.63	1.39	1.47	0.15	0.11	0.17	0.51	0.09	0.09	0.23	0.21	2.58	0.26	0.37	1.06
Fear	34.22	34.74	29.55	29.29	64.39	64.97	49.56	45.40	64.38	61.98	27.30	31.08	30.90	25.77	20.21	23.71
Happiness	13.13	15.43	16.67	17.43	6.53	6.06	16.47	16.98	5.45	6.09	31.68	29.22	19.31	22.70	14.46	15.31
Neutral	33.92	36.11	40.54	39.97	23.12	22.93	26.63	29.04	24.06	25.90	29.15	29.48	35.19	41.07	56.17	51.14
Sadness	3.49	2.69	6.04	6.00	1.14	1.17	2.59	2.94	1.65	1.64	5.99	4.36	5.79	4.34	3.45	3.46

Table 4: Percentage of articles with dominant emotion per news outlet – based on NAWL classification

Emotion	Abortion		Church		Constitution		Elections		EU		Independence March		LGBT		Refugees	
	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP	TVN	TVP
Anger	11.29	18.88	15.06	19.15	11.49	7.69	16.87	18.29	11.15	14.84	8.70	12.79	3.45	20.59	4.84	19.05
Disgust	11.29	16.08	10.84	10.64	7.43	12.09	7.23	10.98	8.46	14.19	6.52	15.12	10.34	26.47	11.29	14.29
Fear	30.65	30.77	29.52	25.53	29.05	42.86	34.94	35.37	27.69	29.68	36.96	38.37	34.48	17.65	33.87	36.05
Happiness	16.13	5.59	17.47	8.51	21.62	13.19	9.64	7.32	26.54	11.61	19.57	16.28	17.24	5.88	19.35	6.12
Neutral	17.74	14.69	15.66	14.89	15.54	13.19	19.28	17.07	16.92	14.19	19.57	8.14	20.69	26.47	18.55	11.56
Sadness	12.90	13.99	11.45	21.28	14.86	10.99	12.05	10.98	9.23	15.48	8.70	9.30	13.79	2.94	12.10	12.93

Table 5: Percentage of articles with dominant emotion per news outlet – based on XLM-RoBERTa Large fine-tuned on XNLI classification

After performing the t-test for all emotions, we noticed that at 5% level of significance, we reject the null hypothesis on no difference between means only for “Abortion” and “Church” categories. For better visibility of this outcome we display the counts of classified examples with the regard to news outlet for “Abortion” category on Figure 1.

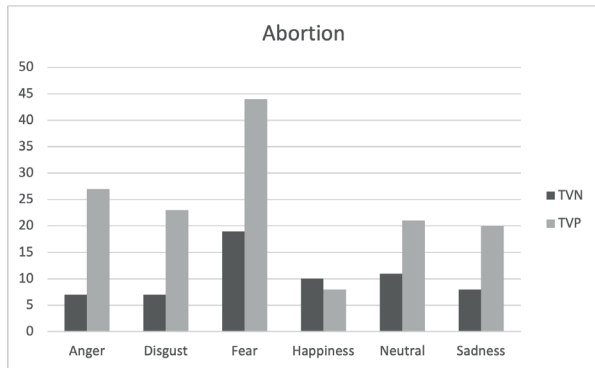


Figure 1: Number of articles classified by RoBERTa-XNLI from category “Abortion”

6. Discussion

Results of error analysis of the second experiment are showed in Table 6. Up to ten words with highest classification probability are displayed (their English translations are presented in brackets). There is a subtle difference in meaning between “Disgust” and “Anger” classes for words like foolishness or absurd. However, five misclassifications which occurred in “Happiness” category related words hold a completely opposite emotional appeal.

Additionally, a study on how to measure controversy is worth taking into consideration. Moreover, the differences in our results are caused by different level classification –

Emotion	Misclassified words
Disgust	głupota, przekleństwo, absurd, tort, malkontent, niezgodny, niewygodny, wścibski, szemrać, zniechęcony (foolishness, curse, absurd, cake, malcontent, inconsistent, uncomfortable, nosy, murmur, discouraged)
Happiness	zadufany, moralista, współczuć, omen, łufcik (complacent, moralist, to sympathize, omen, backlash)

Table 6: Examples of words misclassified by RoBERTa-XNLI

word-level for lexicon-based method and article-level for language model. The length of the articles could have had an influence on results for the former method, in contrast to the latter one. Dictionaries are simpler and faster in use, but have problems with long texts and detection of negations, sarcasm or irony. Also, according to (Sun et al., 2019), in case of long texts which cannot be processed by BERT, we should consider truncating the articles so that only the beginning and the end are left.

7. Conclusion and Future Work

In this work we performed sentiment analysis on a subset of the Polish Online News Corpus which focused on controversial topics. We checked if there are any differences in sentiment of particular topics between news providers. For this task, we compared a lexicon-based and a statistical method of classification. Our work showed that there were no significant differences of emotional appeal on a word-level analysis. However, they were observed in case of results of zero-shot classification between analyzed news outlets for “Abortion” and “Church” categories. In the former topic, three times more words occurring in TVN articles were classified as happy in comparison to TVP. For the latter news category, a classifier showed that TVP had al-

most twice more words describing sadness and at the same time two times less expression of happiness compared with TVN.

Our work is a contribution to quantitative studies on sentiment analysis of Polish news articles. So far, most works were qualitative or based only on short pieces of texts such as tweets or opinions from rating websites, but not on online news. In the future we plan to take into account more controversial topics and analyze the dynamics of emotional load in time by considering publication dates of articles. In case of a lexicon-based approach, we intend to examine overall emotional load of the articles instead of focusing on a dominant sentiment and to run corresponding experiments on other dictionaries available in Polish. Moreover, we aim at classifying the emotional overtone of articles by human annotators. The dataset will be narrowed down to pairs of corresponding news from both media outlets and articles will be chunked into sentences to allow more accurate annotation. This will enable further comparison of zero-shot classification as well as other models results. Additionally, annotators will be asked to state whether selected sentences indicate political leaning. New sentence-level, annotated dataset will be used for research on bias and polarization studies of Polish media.

References

- Aggarwal, Swati, Tushar Sinha, Yash Kukreti, and Siddharth Shikhar, 2020. Media bias detection and bias short term impact assessment. *Array*, 6:100025.
- Baran, Joanna, Michał Kajstura, Maciej Ziolkowski, and Krzysztof Rajda, 2022. Does Twitter know your political views? POLiTweets dataset and semi-automatic method for political leaning discovery. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*.
- Burbach, Laura, Patrick Halbach, Martina Ziefle, and André Calero Valdez, 2020. Opinion formation on the internet: The influence of personality, network structure, and content on sharing messages online. *Frontiers in Artificial Intelligence*, 3.
- Catelli, Rosario, Serena Pelosi, and Massimo Esposito, 2022. Lexicon-based vs. BERT-based sentiment analysis: A comparative study in Italian. *Electronics*, 11:374.
- Cui, Jingfeng, Zhaoxia Wang, Seng Ho, and Erik Cambria, 2023. Survey on sentiment analysis: evolution of research methods and topics. *Artificial intelligence review*:1–42.
- Elejalde, Erick, Leo Ferres, and Eelco Herder, 2018. On the nature of real and perceived bias in the mainstream media. *PLoS ONE*, 13.
- Kocoń, Jan, Maciej Piasecki, and Monika Zaśko-Zielińska, 2017. PIWordNet as a Basis for Large Emotive Lexicons of Polish. In *Human language technologies as a challenge for computer science and linguistics: 8th Language Technology Conference*.
- Machete, Paul and Marita Turpin, 2020. The use of critical thinking to identify fake news: A systematic literature review. *Responsible Design, Implementation and Use of Information and Communication Technology*, 12067:235 – 246.
- Mello, Caio, Gullal Cheema, and Gaurish Thakkar, 2022. Combining sentiment analysis classifiers to explore multilingual news articles covering London 2012 and Rio 2016 Olympics. *International Journal of Digital Humanities*:1–27.
- Mohammad, Saif and Peter Turney, 2013. NRC emotion lexicon. *National Research Council of Canada*:1–234.
- Mäntylä, Mika, Daniel Graziotin, and Miikka Kuutila, 2016. The evolution of sentiment analysis - a review of research topics, venues, and top cited papers. *Computer Science Review*, 27.
- Nandwani, Pansy and Rupali Verma, 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11.
- Pelicon, Andraž, Marko Pranjic, Dragana Miljkovic, Blaž Škrlj, and Senja Pollak, 2020. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10:5993.
- Probierz, Eryka and Adam Galuszka, 2022. Emotion detection based on sentiment analysis: An example of a social robots on short and long texts conversation. *European Research Studies Journal*, XXV:135–144.
- Riegel, Monika, Małgorzata Wierzba, Marek Wypych, Łukasz Żurawski, Katarzyna Jednoróg, Anna Grabowska, and Artur Marchewka, 2015. Nencki affective word list (NAWL): the cultural adaptation of the berlin affective word list–reloaded (BAWL-R) for Polish. *Behavior Research Methods*.
- Shirsat, Vishal S., Rajkumar S. Jagdale, and S. N. Deshmukh, 2017. Document level sentiment analysis from news articles. In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*.
- Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang, 2019. How to fine-tune bert for text classification? In Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu (eds.), *Chinese Computational Linguistics*.
- Szwoch, Joanna, Mateusz Staszko, Rafał Rzepka, and Kenji Araki, 2022. Creation of Polish online news corpus for political polarization studies. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*.
- Tomanek, Krzysztof, 2014. *Sentiment analysis: history and development of the method within CAQDAS (in Polish: Analiza sentymentu: historia i rozwój metody w ramach CAQDAS)*.
- Võ, Melissa, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus Hofmann, and Arthur Jacobs, 2009. The berlin affective word list reloaded (bawl-r). *Behavior research methods*, 41:534–8.
- Zhao, Lingyun, Lin Li, and Xinhao Zheng, 2020. A BERT based sentiment analysis and key entity detection approach for online financial texts. *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*:1233–1238.