Proceedings of the First Workshop on Natural Language Processing for Political Sciences (PoliticalNLP), pages 86–90
Marseille, 24 June 2022

Creation of Polish Online News Corpus for Political Polarization Studies

[1]Joanna Szwoch, [2]Mateusz Staszkow, [3]Rafal Rzepka, [3]Kenji Araki

[1]Graduate School of Information Science and Technology, Hokkaido University,
[2]Mateusz Staszkow Software Development,
[3]Faculty of Information Science and Technology, Hokkaido University
[1,3]Sapporo - Japan, [2] Warsaw - Poland
joannaeleonora.szwoch.u0@elms.hokudai.ac.jp, mateuszstasz@gmail.com
{rzepka, araki}@ist.hokudai.ac.jp

Abstract

In this paper we describe a Polish news corpus as an attempt to create a filtered, organized and representative set of texts coming from contemporary online press articles from two major Polish TV news providers: commercial TVN24 and state-owned TVP Info. The process consists of web scraping, data cleaning and formatting. A random sample was selected from prepared data to perform a classification task. The random forest achieved the best prediction results out of all considered models. We believe that this dataset is a valuable contribution to existing Polish language corpora as online news are considered to be formal and relatively mistake-free, therefore, a reliable source of correct written language, unlike other online platforms such as blogs or social media. Furthermore, to our knowledge, such corpus from this period of time has not been created before. In the future we would like to expand this dataset with articles coming from other online news providers, repeat the classification task on a bigger scale, utilizing other algorithms. Our data analysis outcomes might be a relevant basis to improve research on a political polarization and propaganda techniques in media.

Keywords: Polish language, news corpus, classification, NLP, web scraping

1. Introduction

Nowadays, a piece of information is the most valuable asset. There is a growing problem of distinguishing valuable information from noise and fake news. In Poland it is significantly noticeable during the Russia-Ukraine conflict. One can see inaccuracies regarding the influx of Ukrainian refugees into the European Union, the course of the fighting or the actions of Western countries toward Russia.

In the 21st century we can observe a sociological phenomenon called the filter bubble (Cisek and Krakowska, 2018). Sometimes the same topic is presented in extremely different ways, depending on the news provider. Users tend to visit news sources matching their political attitudes and spend more time on biased content (Garimella et al., 2021). Manipulation is performed with a variety of language techniques and each language requires a tailored approach to detect them.

Polish language cannot be called a typical lesser-resourced language, but compared to others, such as English, German or Russian, it has a significantly smaller base of available corpora. Additionally, vast majority of text resources are paid and not disclosed to the public. The National Corpus of Polish[1] is one of few initiatives which provides a reference corpus containing roughly fifteen hundred million words for free. Sources include literature, newspapers, specialist magazines, transcripts of conversations and Internet texts. However, none of them are online news websites. Moreover, the project was finished in 2012 and has not been updated since. This means that all resources come from year 2011 or older, sometimes even reaching back to 1920s.

Modern languages are very flexible and their use changes constantly. That is why we think that the aforementioned corpora should be expanded by newer sources. Websites providing online news in Polish should therefore be perfect for that as they contain contemporary version of the language in everyday use. Another premise is that they are written in a correct manner, prepared by professional journalists, unlike other online sources such as blogs or social media.

Our dataset was created with the use of two online news websites, run by Polish major TV news providers, namely state-owned TVP Info[2] and commercial TVN24[3]. These are examples of the most watched TV news programs in Poland[4]. Aforementioned websites were scraped and news from years 2019-2021 from different categories were persisted into CSV files. This paper focuses on two tasks - data collection aspect and news outlet classification, which can be treated as a baseline for further experiments. It explains step by step how our corpus was created – we describe web scraping method which we find the most convenient,

[1] http://nkjp.pl/index.php

[2] https://www.tvp.info/

[3] https://tvn24.pl/

[4] https://www.wirtualnemedia.pl/artykul/fakty-lider-ogladalnosci-luty-programy-informacyjne

readable and therefore reproducible. Having created the dataset with scraped articles, we explain how to perform data cleaning to prepare the corpus for modeling. After lemmatization and tokenization, cleaned text is vectorized and classification task is performed with the use of several machine learning models. Unlike most of experiments concerning news articles processing, we trained models to predict which news provider wrote certain article, instead of predicting the article category. People often state that media which do not share their political views or opinions are biased and on the contrary, the ones that they follow are not. In the era of news flowing constantly from different sources it would be beneficial to be able to measure media bias objectively. It is claimed to be possible via data-driven analyses whose results should be free of subjectivity (D'Alonzo and Tegmark, 2021). This topic may be of particular interest in connection with the on-going war in Ukraine as reports about it are presented differently, sometimes to an extreme extend, depending on the source.

We think that this corpus is a good starting point for further analyses of contemporary Polish language. Although, the professional journalists should focus on conveying a clear message, holding subjective information only, we believe that news outlets could be examined whether they show any political bias and what kind of propaganda techniques are being used, if any. However, we want to direct our attention to that matter in our future works, with the use of this data set, possibly extended with more articles from other online news sources[5].

The rest of this paper is structured as follows. In Section 2 we focus on previous works which dealt with the problem of dataset creation, especially news corpora, as well as classification task in NLP. Section 3 describes our dataset, with details of each step of creation process that leads to its final form. In Section 4 we discuss methodology used for our models. Section 5 presents the results of our experiment, whose goal is to find the best way to predict which media outlet created certain piece of news. In Section 6 we summarize conclusions which can be drawn from this article and in the end, we mention our future work plans.

2. Related Works

In the past few years there was plenty of studies which tried to tackle the problem of collecting online text resources in an efficient manner. Researchers from MIT managed to collect over three million articles from 2019 and 2020 from about 100 online media outlets with the use of the open-source Newspaper3k[6] software (D'Alonzo and Tegmark, 2021). Another way to do it is web scraping. This process consists of a few steps: desired websites identification, URLs collection, HTML retrieval, text parsing and finally, persistence (Victoriano et al., 2022). One of the most popular and widely used Python libraries for imitating human alike behavior of entering the URL and retrieving necessary data is Beautiful Soup which automatically obtains data from HTML and XML files (Onyenwe et al., 2021).

Slovak Categorized News Corpus was created in a similar way. It contains words, automatic morphological as well as named entity annotations. It consists of almost five thousand articles, with over one hundred thousand sentences and million and a half of tokens (Hladek et al., 2014).

For text classification task itself, firstly, input data needs to be converted to a computer-readable form. BagofWords and TF-IDF word vectors are two possibilities to handle this task (Qader et al., 2019), (Vimal, 2020). Then, logistic regression can be trained on such data. Except for logit models, other methods include neural networks, support vector machines, random forests, or naive Bayes classifiers (Stein et al., 2020).

3. Corpus Creation

We decided to collect online articles from two Polish major TV news providers, TVP Info and TVN24. Firstly, we tried using Newspaper3k library for this task. Although library documentation states that it handles Polish language, it failed to parse chosen websites correctly and therefore we had to give up on this method. However, one feature that worked properly was listing the subcategories of the main website. In the end we decided to prepare a tailored solution for scraping these resources with Beautiful Soup library in Python, as it was suggested in other works (Onyenwe et al., 2021), (vanden Broucke and Baesens, 2018) (Al Qadi et al., 2019).

3.1. Data Collection

Aforementioned websites were scraped and news from different categories were collected between 1st January 2019 and 31st December 2021.

Data collection cleaning process consisted of the following steps:

- Identifying main pages of news providers
- Listing all contexts (website subpages)
- Web crawling to gather all URLs from designated time span
- Parsing websites to retrieve data from HTML files; first text cleaning with the use of regular expressions to filter markups from retrieved HTMLs
- Saving data to CSV file

The aforementioned process resulted in the collection of articles from two sources in the following amounts and categories presented in Table 1.

[5]Upon request, our dataset can be provided for research purposes.

[6]https://newspaper.readthedocs.io/en/latest/

Category	TVP Info	TVN24
POLAND	36,223	37,511
WORLD	19,982	28,318
SOCIETY	11,484	-
BUSINESS	4,488	12,629
WARSAW	-	12,532
SPORT	3,628	26,289
SCIENCE	2,297	108
CULTURE	1,698	-
MISCELLANEOUS	1,399	-
WEATHER	495	10,558
POLITICS	-	513
ENTERTAINMENT	-	69
TOTAL	81,694	128,527

Table 1: Number of articles within each category

Data is not evenly distributed and some categories existed only in one website, but not in the other one. Datasets have the following structure as presented in Table 2.

Variable	TVP Info	TVN24
`url`	✓	✓
`magazine_title`	✓	✓
`website_category`	✓	✓
`title`	✓	✓
`description`	✓	✓
`authors`	✓	✓
`article`	✓	✓
`pub_time`	✓	✓
`mod_time`	✓	-
`hash_tags`	✓	-

Table 2: Extracted data

TVP Info dataset has two additional columns when compared to TVN24, namely modification time and hash tags. We decided to include them as they might be interesting to be examined in the future for other purposes. TVN24 dataset did not have any information regarding the modification time of the article and tags appeared only recently in their articles, therefore these columns were not added to this dataset.

3.2. Dataset Cleaning

Dataset cleaning process consisted of 5 steps in the following order:

- **Duplicates removal** - some articles were repeated during the web crawl.
- **Removing repetitions from retrieved text** - in some cases, description of the article appeared also in the article text which was not desired.
- **Filtering ad words** - most of the articles consisted of phrases which encouraged the reader to watch a related video or read an article whose topic is connected.
- **Deleting special characters such as punctuation, double spaces or tabulation as well as numbers** - regular expressions were used to eliminate all unnecessary elements from this group.
- **Stop words removal** - we used *stop-words* Python library[7] and a set listed by user *bieli* on GitHub[8] to create an extended collection of Polish stop words, as part of them were not included in *SpaCy* library.

As a result, we obtained fairly clean 197,606 articles from both TVP and TVN, consisting of 4,042,638 sentences and 44,528,641 tokens.

4. Methodology

In this Section all methods which were used to perform text classification task are briefly explained.

4.1. Dataset Preparation

Firstly, cleaned dataset has to undergo a few more processes before it is eventually used as an input to train classification models, namely:

- **Lemmatization** - *SpacyPL* handled this for Polish language, using a lemma dictionary imported from Morfeusz morphological analyzer[9].
- **Tokenization** - *nltk* Python library was used for this task.

As the website category groups were unevenly distributed, we decided to take 2,000 randomly selected articles both from TVN24 and TVP Info from the following four most numerous categories: WORLD, POLAND, BUSINESS and SPORTS. Eventually, we trained our models on a reduced subset of 16,000 online news. Dataset was then divided into two sets - training set which consists of 12,000 records and test set that has remaining 4,000 tuples.

4.2. Feature Extraction

Although classifying of text is not an easy task to be performed by computers, it can be done if input data is converted into a numerical representation.

- **Bag of Words (BoW)** - this method is considered to be simpler both computationally and conceptually than other methods. It is assumed that it could record higher performance scores on common used benchmarks of text (Qader et al., 2019).

[7] `https://github.com/Alir3z4/python-stop-words`
[8] `https://github.com/bieli/stopwords`
[9] `http://morfeusz.sgjp.pl/`

- **TF-IDF** - word vectors are also able to help with converting characters into a format that is processable by a computer (Vimal, 2020).

Both methods were implemented with the use of *Scikit-learn* library.

4.3. Algorithms

Based on the suggestions from previous works, we trained following four machine learning models:

- **Logistic regression** - commonly used for NLP classification tasks such as fake news detection (Yu et al., 2021)
- **Random forest** - recommended for news articles classification task, along with N-gram textual features (Liparas et al., 2014)
- **Support Vector Machine** - better results with high dimension data like large volumes of text, comparing with Neural Networks or Naive Bayes methods (Shahi and Pant, 2018)
- **Naive Bayes** - high accuracy in online news category classification task (Khine and Nwet, 2016)

Scikit-learn Python library allows us to use already built-in functions which perform all the calculations of the aforementioned supervised learning methods[10].

4.4. Model Quality Measurements

In order to check the performance of trained models, we calculated two of the standard metrics, namely **Accuracy** and **F1 Score** (Blagec et al., 2020). We also use *Scikit-learn* library for retrieval of these statistics.

5. Results

Having trained four different models with two possible types of feature extraction, we came with the following results as shown in Table 3.

	Feature Representation			
	BoW		TF-IDF	
Model	Accuracy	F-score	Accuracy	F-score
SVM	0.8668	0.8713	0.8598	0.8629
Random forest	0.8703	0.8798	**0.8745**	**0.8829**
Logistic Regression	0.8700	0.8729	0.7228	0.7790
Naive Bayes	0.8048	0.8013	0.7640	0.7627

Table 3: Performance measurements of models

The best results were achieved by random forest with TF-IDF vectorization method. Accuracy was 87.45% and F-score was equal to 88.29%. The worst results were obtained by logistic regression which also had features vectorized with TF-IDF method. Accuracy and F-score were equal 72.28% and 77.90% accordingly and these were the lowest scores among other models.

[10] https://scikit-learn.org/stable/

6. Discussion

TVN24 and TVP Info are two of the most watched news providers in Poland. Nonetheless, the way the information is conveyed by both is considered to be often much different from each other (Klepka, 2017), (Weglińska et al., 2021), (Batorowska et al., 2019). We believe it would be interesting to supplement social studies with machine learning techniques for measuring political bias. Another step to be considered is error analysis to check which words were most often confused when predicting the news outlet.

Our showcase study illustrated that there is still room for an improvement as others achieved over 90% accuracy scores in similar classification problems such as fake news detection (Alenezi and Alqenaei, 2021) or news categorization (Bracewell et al., 2009) as well as news topic analysis (Minaee et al., 2021).

7. Conclusions and Future Works

We introduced a collection of online news from two TV news broadcasters as the first step of building a full-fledged corpus containing modern, journalist-created and hopefully correct sample of Polish language. We showed that this corpus can be used for such a task like news outlet classification based on news article contents. Out of four trained models, random forest with TF-IDF vectorization achieved the highest accuracy and F-score metrics. Our next step is to compare these results with a few more prediction methods including Convolutional Neural Networks or Polbert language model (Kłeczek, 2020).

American news outlets have already been analyzed in terms of political polarization problem (D'Alonzo and Tegmark, 2021). Our dataset is unique and we hope our contribution will become a base for other tasks, such as news category prediction or also sentiment analysis in Polish language. With this corpus of distinctly different sources we also look forward to encouraging researchers to use it to study polarization of Polish society and to develop methods (e.g. for automatic summarization) freeing information of political biases and possible manipulation. In the future, we want to extend the dataset with articles coming from other online news outlets from the same period of time to study the full scale of political gradation of Polish media. It would allow us to focus on further research regarding political bias and detection of propaganda techniques.

8. Bibliographical References

Al Qadi, L., El Rifai, H., Obaid, S., and Elnagar, A. (2019). Arabic text classification of news articles using classical supervised classifiers. pages 1–6, 10.

Alenezi, M. and Alqenaei, Z. (2021). Machine learning in detecting covid-19 misinformation on twitter. *Future Internet*, 13:244, 09.

Batorowska, H., Wasiuta, O., and Klepka, R. (2019). *Media jako instrument wpływu informacyjnego i manipulacji społeczeństwem*. 02.

Blagec, K., Dorffner, G., Moradi, M., and Samwald, M. (2020). A critical analysis of metrics used for measuring progress in artificial intelligence. *CoRR*, abs/2008.02577.

Bracewell, D., Yan, J., Ren, F., and Kuroiwa, S. (2009). Category classification and topic discovery of japanese and english news articles. *Electr. Notes Theor. Comput. Sci.*, 225:51–65, 01.

Cisek, S. and Krakowska, M. (2018). The filter bubble: a perspective for information behaviour research, 10.

D'Alonzo, S. and Tegmark, M. (2021). Machine-learning media bias. *CoRR*, abs/2109.00024.

Garimella, K., Smith, T., Weiss, R., and West, R. (2021). Political polarization in online news consumption, 04.

Hladek, D., Stas, J., and Juhar, J. (2014). The Slovak categorized news corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1705–1708, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Khine, A. H. and Nwet, K. T. (2016). Automatic myanmar news classification using naive bayes classifier.

Klepka, R., (2017). *Ewolucja Wiadomości TVP1: od medialnej stronniczości do propagandy politycznej?*, pages 244–253. 03.

Kłeczek, D. (2020). Polbert: Attacking Polish NLP tasks with transformers. In Maciej Ogrodniczuk et al., editors, *Proceedings of the PolEval 2020 Workshop*. Institute of Computer Science, Polish Academy of Sciences.

Liparas, D., HaCohen-Kerner, Y., Moumtzidou, A., Vrochidis, S., and Kompatsiaris, I. (2014). News articles classification using random forests and weighted multimodal features. volume 8849, 11.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40.

Onyenwe, I. E., Onyedinma, E. G., Nwafor, C. A., and Agbata, O. (2021). Developing products update-alert system for e-commerce websites users using HTML data and web scraping technique. *CoRR*, abs/2109.00656.

Qader, W. A., Ameen, M. M., and Ahmed, B. I. (2019). An overview of bag of words; importance, implementation, applications, and challenges. In *2019 International Engineering Conference (IEC)*, pages 200–204.

Shahi, T. B. and Pant, A. K. (2018). Nepali news classification using naïve bayes, support vector machines and neural networks. In *2018 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–5.

Stein, A., Weerasinghe, J., Mancoridis, S., and Greenstadt, R. (2020). News article text classification and summary for authors and topics. pages 1–12, 11.

vanden Broucke, S. and Baesens, B., (2018). *Examples*, pages 197–298. Apress, Berkeley, CA.

Victoriano, J., Pulumbarit, J., and Lacatan, L. (2022). Data analysis of BulSU faculty research engagement based on Google Scholar data using web data scrapping technique. *International Journal of Computing Sciences Research*, 26:1–12, 01.

Vimal, B. (2020). Application of logistic regression in natural language processing. *International Journal of Engineering Research and*, V9, 06.

Weglińska, A., Szurmiński, , and Wasicka-Sroczyńska, M. (2021). Politicization as a factor of shaping news in the public service media : A case study on public television in poland polityzacja jako czynnik w kształtowaniu przekazu medialnego w tvp sa - studium przypadku. *Athenaeum Polskie Studia Politologiczne*, 72:29–51, 12.

Yu, P., Cui, V., and Guan, J. (2021). Text classification by using natural language processing. *Journal of Physics: Conference Series*, 1802:042010, 03.