A Comparison of Text Representation Methods for Predicting Political Views of Social Media Users

Anna Glazkova[0000-0001-8409-6457]

University of Tyumen, 6, Volodarskogo Str., Tyumen, 625003, Russian Federation
a.v.glazkova@utmn.ru

Abstract. The paper focuses on the task of predicting political views of social media users. The aim of this study is to identify the most effective method for representation textual information from user profile. We compared several text representation methods, including a bag of words modeling, averaged word2vec embeddings, Sentence Transformers representation, and text representations obtained with three BERT-based models, such as Multilingual BERT, Slavic-BERT, and RuBERT. We conducted our experiments on the dataset of VKontakte users' data collected with VK API. We evaluated the effectiveness of binary classification for the pages of users with radical political views, including ultraconservatives, communists, and libertarians, and users who are indifferent to politics. Further, we compared the impact of various text representations for distinguishing users belonging to different radical political movements, such as communists vs. libertarians, libertarians vs. ultraconservatives, ultraconservatives vs. communists. Best results were predictably shown by BERT-based models. Moreover, in each task, the best result was achieved by different models.

Keywords: Social Media, Political Preferences, Text Representation, BERT, VKontakte, Word Embeddings.

1 Introduction

Social media analysis is one of the key tasks of natural language processing and information retrieval. The aim of the analysis of social media is to develop powerful methods and algorithms which extract relevant information from a large volume of social network data [6]. Social media data processing is a significant part of various natural language processing systems, such as social media monitoring and fact checking [8; 13], interest discovery [27], health care [12; 21], business intelligence [31], and security [29; 33].

The widespread use of social network for social and political communication creates many opportunities to monitor the political views of large numbers of people in real time [4]. Thus, social network profiles contain a lot of valuable information that can be used as a material for sociological research or as a tool for political influence [9].

In this paper, we perform a comparison of text representation methods for predicting political views of social media users based on textual data posted on their personal profiles. We consider the following types of text representations: a) a bag-of-words representation [11]; b) averaged Word2vec embeddings [22]; c) text representations from BERT [5], including embedding from Multilingual BERT, RuBERT [16], SlavicBERT [2], and text representation using Sentence Transformers [26]. We conduct our experiments on the dataset collected using VKontakte API* and use textual information from personal profiles of social media users to predict their political preferences.

The paper is organized as follows. Section 2 gives a brief description of text representation methods used in this work. Section 3 is concerned with the dataset used for this study. Section 4 presents our experiments and results. Section 5 is a conclusion, and Section 6 contains acknowledgements.

2 Text Representation Methods

In this section we describe the methods of text representation we used in our work.

Bag of words. The bag of words (BoW) [11] model is a classical approach to text representation for machine learning algorithms. This model describes the occurrence of words within a document. The text is presented as a token counts matrix. The BoW model is widely used in different natural language processing tasks. Nevertheless, it suffers from some shortcomings, such as sparsity and word order ignoring.

Word2vec. The idea of word2vec [22] is based on the assumption that the meaning of a word is affected by the words around it. This statement follows distributional hypothesis [11]. Word embeddings assign a real-valued vector for each word and represent the word by the vector. The averaged word embeddings are calculated by summing of all words' vectors in a text and dividing the sum by the text length.

BERT (Bidirectional Encoder Representations from Transformers). This machine learning technique [5] based on transformer neural architectures presents state-of-the-art results in a wide variety of natural language processing tasks, including text classification, opinion mining, and others. BERT's key innovation is a bidirectional training which helps to consider both left and right contexts.

In our work, we evaluated the following models:

- Multilingual BERT [5], a pretrained model by Google on the top 104 languages with the largest Wikipedia using a masked language modeling (MLM) objective.
- RuBERT [16], a model by DeepPavlov [3] trained on the Russian part of Wikipedia and news data. Multilingual BERT was used as an initialization for RuBERT.
- SlavicBert [2], a model trained on Russian News and four Wikipedias: Bulgarian, Czech, Polish, and Russian. Multilingual BERT was also used as an initialization for SlavicBERT.
- Text representation obtained with Sentence Transformers [26], a framework that provides a method to compute dense vector representations for sentences and para-

* https://vk.com/dev/first_guide

graphs based on BERT-based networks. In our experiments, we used a multilingual knowledge distilled version of multilingual universal sentence encoder [32] trained for the task of similar text detection. This multilingual knowledge distilled version supports 50+ languages.

3 Dataset

To conduct our experiments, we collected data using VKontakte API. The personal profile of the VKontakte user contains a list of text and categorical fields. The user can fill in these fields with his personal information. In particular, he can indicate his political views by choosing one of 9 possible options, including moderate, conservative, liberal, socialist, monarchist, ultraconservative, libertarian, communist, and indifferent.

The aim of this study is to compare the impact of different text representations for predicting political views of social network users. For example purposes, we decided to evaluate the effectiveness of binary classification of users identified radical political views, such as ultraconservatives, communists, and libertarians, and users who are indifferent to politics. Moreover, we compared the impact of text representations for distinguishing users belonging to different radical political movements.

For this purpose, we downloaded textual information from users' personal profiles. This information is contained in text fields that are filled in by users in a free form. These fields include descriptions of the user's activities, favorite music, movies, TV shows, games, sources of inspiration, and the user's worldview. Text fields are mostly filled in Russian, since VKontakte is especially popular in post-Soviet countries. However, there are a large number of texts in other languages, for example English and various Slavic languages [1; 7; 15].

We combined text from all text fields of the profile and selected only those users whose texts have a total lenght of at least 10 words. Further, for our experiments, we selected users who indicated ultraconservative, communist, libertarian, or indifferent political views. Table 1 shows the main characteristics of our data.

Table 1. Corpus description.

Political views	Number of texts	Avg length (number of words)	Avg length (number of symbols)
Communist	299	31.1	260.06
Ultraconservative	240	27.66	222.91
Libertarian	116	46.51	392.5
Indifferent	799	38.63	316.09

4 Results and Discussion

We conducted our experiments on Google Colab Pro[†] (CPU: Intel(R) Xeon(R) CPU @ 2.20GHz; RAM: 25.51 GB; GPU: Tesla P100-PCIE-16GB with CUDA 10.1).

Each BERT-based model (mBERT, SlavicBERT, and RuBERT) was fine-tuned on the training set for 2 epochs. We used random seeds to fine-tune pretrained language models and made attempts to combine them with other parameters. The models are optimised using AdamW [20] with a learning rate of 2e-5, epsilon of 1e-8, max sequence length of 128 tokens, and a batch size of 32. We implemented our models using Pytorch [23] and Huggingface's Transformers [30] libraries.

We utilized word2vec embeddings provided by RusVectores[‡] [17]. The model was trained on Russian Wikipedia[§] texts collected in 2018. The vector size is 300. The model was loaded and processed with Gensim [25]. Finally, we implemented the BoW representation using Scikit-learn Python library [24].

To reduce the class imbalance, we used a random oversampling technique implemented with Imbalanced-learn Python library [18]. Random oversampling involves supplementing the training data with multiple copies of some minority class examples and can be a fast and effective solution for the problem of class imbalance [10, 28].

We applied a Linear Support Vector Machine (Linear SVC) as a classifier with a tolerance for stopping criteria equal to 1e-5. The classifier takes as an input various text representations sequentially. We used the weighted F1-score as an evaluation metric. The classifier was implemented with Scikit-learn [24]. We splitted our data to train and test datasets in a 80-20 ratio and performed a 3-fold cross-validation. To preprocess data, we used Pymorphy2 [14] and NLTK [19].

The results are presented in Table 2.

Table 2. Results (weighted F1-score, %).

Text representation	Indifferent - others	Communists - libertarians	Libertarians - ultraconservatives	Ultraconservatives - communists	Avg F1-score
BoW	56.74	65.91	78.65	52.04	63.34
Word2vec	63.61	64.42	70.05	53.19	62.82
Sentence Transformers	67.43	64.17	**80.97**	51.68	66.06
mBERT	65.41	**65.97**	79.01	52.23	65.66
RuBERT	66.12	64.8	78.61	**57.18**	66.68
SlavicBERT	**67.95**	65.27	77.05	57.17	**66.86**

[†] https://colab.research.google.com/
[‡] https://rusvectores.org/en/
[§] https://ru.wikipedia.org/

As can be seen from the table above, the best results in all tasks were obtained using BERT-based models. SlavicBERT archived 67.95% of F1-score on the indifferent vs. others task. The classifier trained on mBERT embeddings showed 65.97 for the communists vs. libertarians task. For the libertarians vs. ultraconservatives task, the best result was shown with Sentence Transformers embeddings (80.97%). The highest result for the ultraconservatives vs. communists task was achieved by RuBERT (57.18%). The best averaged result was shown by SlavicBERT (66.86%).

It can be seen from the data in Table 2 that the results for the ultraconservatives vs. communists task are lower than for other tasks. At the same time, all classifiers show their best results for the libertarians vs. conservatives task. This fact can be useful when studying the interests of social groups with different political views.

5 Conclusion

In this study, we compared several methods to represent textual data from users' profiles on social networks. The best results were obtained with BERT-based models, which now show the state-of-the-art achievements in many natural language processing tasks. In our further work, we plan to explore various ways of representing different types of features for predicting political views in social media.

6 Acknowledgments

The reported study was funded by RFBR and EISR, project number 20-011-32031.

References

1. Anisimova, O., Vasylenko, V., Fedushko, S.: Social networks as a tool for a higher education institution image creation. arXiv preprint arXiv:1909.01678 (2019).
2. Arkhipov, M. et al.: Tuning multilingual transformers for language-specific named entity recognition. In: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, 89-93 (2019).
3. Burtsev, M. et al.: Deeppavlov: Open-source library for dialogue systems. In: Proceedings of ACL 2018, System Demonstrations, 122-127 (2018).
4. Conover, M. et al. Predicting the Political Alignment of Twitter Users. In: ASSAT/SocialCom 2011, 192-199, IEEE, Boston (2011).
5. Devlin, J. et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
6. Farzindar, A., Inkpen. D.: Natural language processing for social media. Synthesis Lectures on Human Language Technologies, 2(8), 1–166 (2015).
7. Feshchenko, A. V. et al.: Analysis of user profiles in social networks to search for promising entrants. In: INTED2017: 11th International Technology, Education and Development Conference, 5188-5194 (2017).

8. Glazkova, A., Glazkov, M., Trifonov, T. g2tmn at Constraint@ AAAI2021: Exploiting CT-BERT and Ensembling Learning for COVID-19 Fake News Detection. arXiv preprint arXiv:2012.11967 (2020).
9. Glazkova, A., Sokova, Z., Kruzhinov, V.: Predicting Political Views in Social Media: VKontakte as a case study. https://osf.io/preprints/27ku6/, last accessed 2020/12/22.
10. Glazkova, A.: A Comparison of Synthetic Oversampling Methods for Multi-class Text Classification. arXiv preprint arXiv:2008.04636 (2020).
11. Harris, Z. S.: Distributional structure. Word 10(2-3), 146-162 (1954).
12. Jiang, L., Yang, C. C.: User recommendation in healthcare social media by assessing user similarity in heterogeneous network. Artificial intelligence in medicine 81, 63-77 (2017).
13. Kim, J. H. et al.: Understanding Social Media Monitoring and Online Rumors. Journal of Computer Information System, 1–13 (2020).
14. Korobov, M.: Morphological analyzer and generator for Russian and Ukrainian languages. In: International Conference on Analysis of Images, Social Networks and Texts, 320-332 (2015).
15. Krylova, I. et al.: Languages of Russia: Using social networks to collect texts. In: Russian summer school in information retrieval, 179-185 (2015).
16. Kuratov, Y., Arkhipov, M.: Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. arXiv preprint arXiv:1905.07213 (2019).
17. Kutuzov, A., Kuzmenko, E.: WebVectors: a toolkit for building web interfaces for vector semantic models. In: International Conference on Analysis of Images, Social Networks and Texts, LNCS, 155–161, Springer, Cham (2016).
18. Lemaître, G., Nogueira, F., Aridas, C. K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research, 18(1), 559-563 (2017).
19. Loper, E., Bird, S.: NLTK: the natural language toolkit. arXiv preprint cs/0205028 (2002).
20. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101 (2017).
21. Lu, Y. et al.: Understanding health care social media use from different stakeholder perspectives: a content analysis of an online health community. Journal of medical Internet research, 4(19), e109 (2017).
22. Mikolov, T. et al.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26, 3111-3119 (2013).
23. Paszke, A. et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 8026-8037 (2019).
24. Pedregosa, F. et al.: Scikit-learn: Machine learning in Python. The Journal of machine Learning research, 12, 2825-2830 (2011).
25. Řehůřek, R., Sojka, P.: Gensim—statistical semantics in python. Retrieved from genism.org (2011).
26. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3973-3983 (2019).
27. Shahzad B. et al.: Discovery and classification of user interests on social media. Information Discovery and Delivery (2017).
28. Suh, Y. et al.: A comparison of oversampling methods on imbalanced topic classification of Korean news articles. Journal of Cognitive Science, 18(4), 391-437 (2017).
29. Walsh, J. P.: Social media and border security: Twitter use by migration policing agencies. Policing and Society, 10(30), 1138-1156 (2020).

30. Wolf, T. et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38--45 (2020).
31. Xu, X. et al.: Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors. International Journal of information management, 6(37), 673-683 (2017).
32. Yang, Y. et al.: Multilingual universal sentence encoder for semantic retrieval. arXiv preprint arXiv:1907.04307 (2019).
33. Zhang, Z., Gupta, B. B.: Social media security and trustworthiness: overview and new direction. Future Generation Computer Systems, 86, 914-925 (2018).