Advances in Social Science, Education and Humanities Research, volume 617
Proceedings of the International Conference on Social Science, Psychology and Legal Regulation (SPL 2021)

Sentiment Analysis Based on the BERT Model: Attitudes Towards Politicians Using Media Data

Svitlana Salnikova * [0000-0001-6368-9480], Roman Kyrychenko [0000-0002-1610-4352]

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
* *sv.salnikova@gmail.com*

ABSTRACT
The latest analysis methods of sentiments borrowed from computational linguistics are relevant in the age of big data, which is difficult to process through traditional content analysis. These methods have made it possible to analyze information over a long period, which allows us to trace the dynamics of the relationship to a particular object over time and large-scale comparative studies of texts. The authors demonstrate the applicability of sentiment analysis based on transformer models to the study of the temporal model of attitudes towards well-known politicians (2001-2021) on the example of text analysis of multilingual online publications. To do this, the authors used the targeted-BERT method for automated directed analysis of sentiments, obtained quality indicators F1-score 0.799 and 0.741 for Ukrainian and Russian models, respectively. The authors tested the dependence of mediatization of politicians on the country's political hierarchy, confirmed hypotheses about the attitude to their power (more significant criticism of the Ukrainian media and gradual loyalty to the Russian media) and foreign politicians (dominance of negative tone in both media with a growing trend for Ukrainian media).

Keywords: *computational text analysis, targeted-BERT, tone dynamics, neural network, model training.*

1. INTRODUCTION

In recent decades, most textual information has become available in digital format. In particular, all media have the official status of online publications. Thus, large amounts of information have accumulated, which on the one hand are a source of primary information for researchers, and on the other - require automated approaches to its analysis. Classical approaches applied on their own are no longer enough. Sentiment analysis (SA) is a class of computational methods that automatically retrieves and summarizes opinions about such a vast amount of data that the researcher alone cannot process.

Automated word processing has also made significant progress. Ten years ago, the standard in text analysis was the bag-of-words approach [1], which five years later changed to the approach using neural networks based on recurrent layers (LSTM, GRU) [2], now considered the standard models with transformer architecture [3]. These changes brought a technological complication of approaches and an increase in the ability of machines to understand language, including based on symbols other than Latin.

2. RESEARCH METHODOLOGY

Sentiment analysis, SA (or mood analysis, tone analysis), is a multidisciplinary field of research that deals with the analysis of moods, evaluations, emotions, and thoughts of people about different subjects (entities) [4], such as products, services, individuals, companies, organizations, events and topics. It includes several areas, such as natural language processing (NLP), computational linguistics, information retrieval, machine learning, and artificial intelligence. It is a set of computational methods or techniques [5] and those based on NLP, which can be used to extract subjective information in a particular text.

An essential property of SA is the fixation of polar estimates (negative-positive); its operationalization is to find ways to encode such estimates. From the point of view of machine learning, SA is the task of binary classification into negative and positive classes.

The **purpose** of the article is to analyze the dynamics of the tones of Ukrainian and Russian media reports about famous politicians using new methods of sentimental analysis of texts.

The **main hypotheses** of the study: (1) Ukrainian media will be more critical of its government than Russian to its own; (2) Russian media will be more negative about foreign politicians than Ukrainian; (3) Over time, the Russian media became more loyal to their government. We will also examine the “mediatization of politicians”, particularly in assessing politicians loyal and disloyal to the government [6].

2.1. Targeted sentiment analysis model architecture

Classical sentiment analysis is not aspect-oriented, but there can be many different sentiments in one text, focused on different people, objects, events, etc. That is why the development of methods of aspect-based sentiment classification (ABSC) is now very relevant. In this article, we use the approach of Chinese developers, whose idea was to modify the popular language model Bidirectional encoder for transformers (BERT) to work with local and global contexts [7]. The model consists of 2 blocks: the 1st is responsible for the local context, the 2nd - for the global. Each block is a standard BERT model with layers of embeddings (vector representations) at the input (positional embedding and embedding for tokens) and 12 layers of encoders, to which a layer of dynamic content mask is added in the case of the first block. Then the processing results of these blocks are combined in one layer through the mechanism of self-attention and using it already from the combined result the sentiment forecast for pair text/aspect is received.

A key element of this architecture is the CDM, a dynamic content mask that is a mechanism for denoting words (tokens) that are important for defining sentiment about a particular object in the text. It is calculated on the basis of SRD - relative semantic distance. When learning and deriving model predictions, you need to set this parameter, its purpose is to indicate how many words around the target word to take into account as a local context.

The authors tested the quality of the model on three datasets. The F1-score of this model ranged from 75.78 (on the Twitter dataset) to 81.74 (on the restaurant review dataset). Previous technologies to solve this problem, such as BERT-SPC, RAM, TD-LSTM, and others, did not show a result better than 78.79. Therefore, the choice of this model is logical and on the grounds of its better accuracy, and on the grounds of the logic of architecture, which considers two types of context.

2.2. Training approach

The training of the model of tonal analysis for the Russian and Ukrainian languages took place in two stages: the training of a separate model for the Ukrainian language and the training of a separate model for the Russian language. To train the Ukrainian-language model, we used our marked dataset for 7231 news and Twitter messages with the specified tone and the person in respect of whom such a tone was used. For the Russian-language model, we took 13,335 news and Twitter messages that were similarly tagged.

We have randomly divided datasets then so that approximately 90% of the observations went to the training sample (the model learned to identify patterns between text and sentiment) and 10% to the test (we later tested the quality of the trained model). Thus, we selected 6,500 texts for training the Ukrainian-language model and 12,000 for the Russian-language model.

We have used the transfer learning approach to train the models. This is a traditional way of training neural networks, in which we do not train the weight of the neural network from scratch but take as a basis a ready-made model and train its weight for a specific task (in our case, it is a classification of sentiments). The scales of the finished models using the LCF-BERT approach will be loaded into blocks of local and global contexts. Other layers of the neural network will optimize weights from randomly initialized values.

We used two models for training: the Ukrainian-language youscan/ukr-roberta-base (an extended variation of the BERT model, known as Roberta (12 layers)) [8], and the Russian-language DeepPavlov/rubert-base-cased [9].

The trained models achieved F1-score indicators of 0.799 for Ukrainian and 0.741 for Russian, which generally corresponds to the declared quality indicators of the LCF-BERT model (where the authors tested the quality in English texts). Therefore, we can use this model for automated sentiment analysis.

We collected data for analysis through the rvest library in R, and trained models using Pythons library transformers. Further analysis of the obtained results of sentiment analysis we performed in the programming environment R.

2.3. Sample

To study the dynamics of tonalities concerning key politicians, took the media outlets of Ukrainian Pravda (UP) and Komsomolskaya Pravda (KP) as typical representatives of the Ukrainian and Russian media, respectively. We have united these publications by the same year of foundation - 2000, hence analyzing all texts from 2001 to 2021. The logic of the selection is to include politicians who have been the focus of media attention for a long time in the field of analysis. This will make it possible to identify the dynamics of judgments.

3. RESEARCH RESULTS

We have collected a total of 316072 news pieces from UP and 146027 news pieces from KP for analysis.

The sample is complete, but we have not selected some of the news due to technical failures during the machine collection of materials (NA = 21 and 2 for UP and KP, respectively). There are many more mentions of politicians in the UP than in the KP (Table 1). The reason may be that this publication positions itself as political, while the KP is gradually increasing its tabloid orientation.

Table 1. Distribution of news by years

Year	2001	2002	2003	2004	2005	2006	2007
UP	4497	6085	4969	8395	9923	11748	13995
KP	464	657	1347	3161	3906	4221	5438
Year	2008	2009	2010	2011	2012	2013	2014
UP	14930	15627	16538	17171	16599	15898	24045
KP	7687	7761	8391	9712	9792	9203	9098
Year	2015	2016	2017	2018	2019	2020	2021
UP	25790	23648	21286	20607	19093	23592	1615
KP	8658	9338	10238	10050	10250	11847	4790

Among the politicians mentioned frequently on both resources over the past 20 years (Table 2) are the presidents of Ukraine, the United States, and the Russian Federation, some prime ministers, the German chancellor, and a prominent Russian opposition figure. In general, there is a "logic of the political system", which involves the coverage of politicians by journalists per the political hierarchy of the country [10], regardless of the bias of politicians [11].

Table 2. Distribution of news among politicians

Politician	UP	KP	Politician	UP	KP
Bush G.	1396	443	Trump D.	5520	497
Kuchma L.	12585	72	Turchynov O.	5066	52
Medvedev M.	2661	1790	Tymoshenko Y.	30992	270
Merkel A.	2228	226	Yanukovych V.	33102	226
Navalny A.	751	210	Yatsenyuk A.	8555	83
Obama B.	1876	570	Yushchenko V.	28166	157
Poroshenko P.	20025	421	Zelensky V.	9225	478
Putin V.	14429	4377			

The graphs of Figure 1 clearly show that the average rate of negative connotations in reports about Putin and Medvedev every year becomes smaller. Thus, until 2010, the tone of reports about Putin in the UP was less negative than in the KP. Since then, every year, the tone between the publications is more noticeable, and the Ukrainian

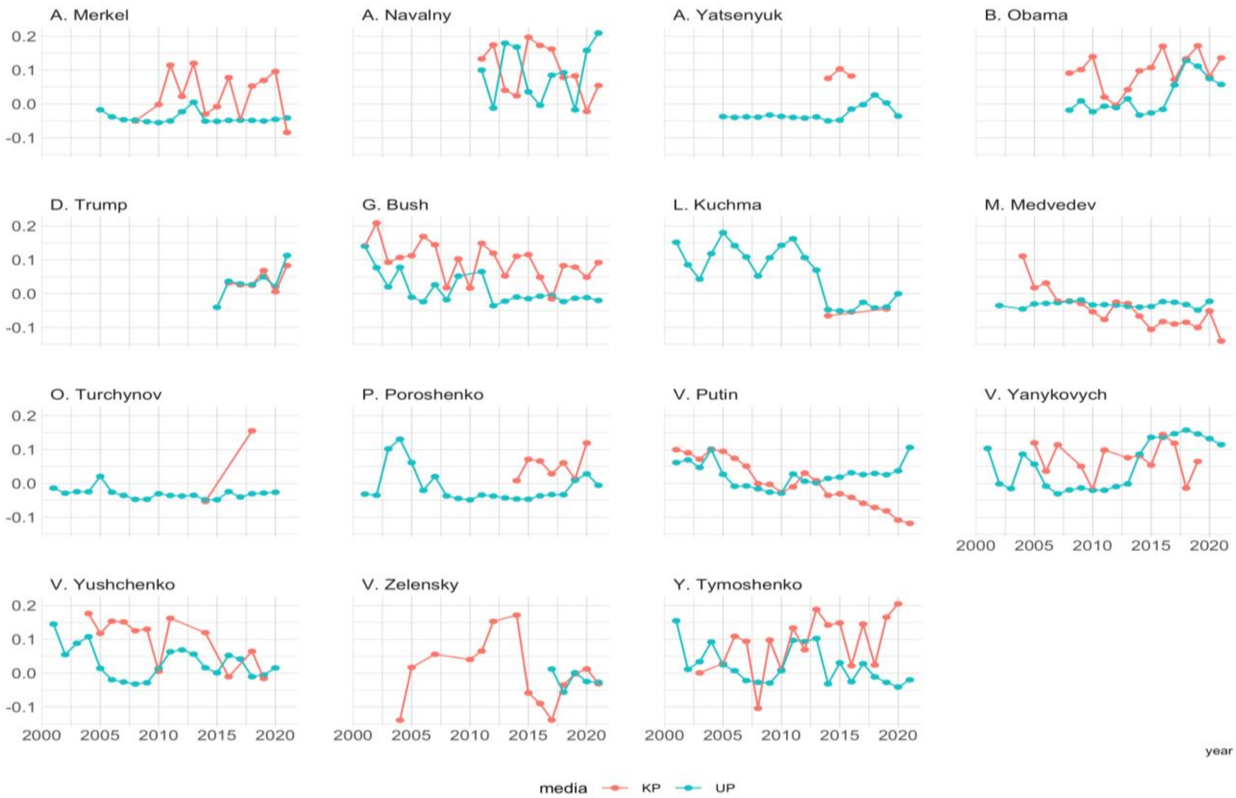

Figure 1 Dynamics of tonality in reports about politicians*, 2001–2021

* Scale: average level of negative attitude, 2000 is considered as zero

edition tends to give a more negative assessment of Putin [12].

The assessment of power in the Ukrainian and Russian media was identical until 2009; afterward, the UP evaluates Ukrainian Presidents less loyally than the KP evaluates Russian Presidents. This difference in estimates increased after 2014 (Figure 2). Throughout the years, the Ukrainian media has been more critical of its government than the Russian apparently because of the UP's opposition to Kuchma's government (not least because of his alleged involvement in the assassination of the publication's founder) and Yanukovych. However, the criticism of the UP is not tied to specific persons in power. In general, the level of positivity of tones in the reports about the Russian authorities in the Russian edition has increased by 20% in 20 years; It is important to note that the editor of KP was Vladimir Putin's proxy in the 2018 presidential election.

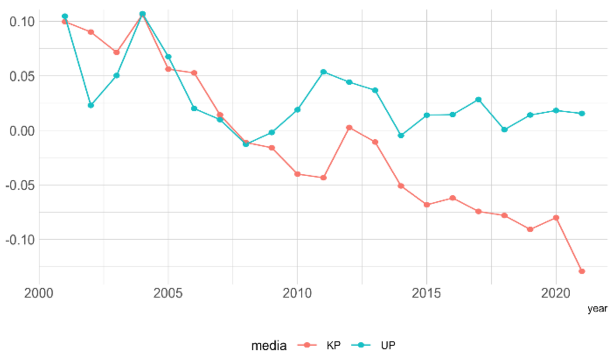

Figure 2 Assessment of power*, 2001–2021

* Scale: average level of negative attitude, 2000 is considered as zero

According to incomplete data in 2021, the average negative assessment of foreign politicians by Ukrainian publications is higher than that of Russian ones. There is also a noticeable tendency for Russian media reports to become less negative, while the Ukrainian media is characterized by the opposite trend (Figure 3).

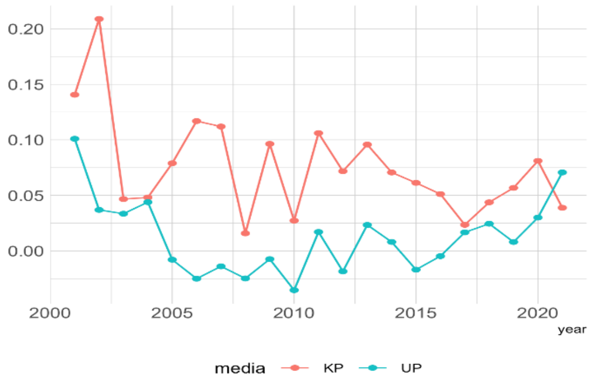

Figure 3 Assessment of foreign politicians*, 2001–2021

* Scale: average level of negative attitude, 2000 is considered as zero

4. DISCUSSION

Within the research framework, the main issues for discussion are related to automated word processing. For most researchers, classical machine learning algorithms remain relevant [13], sometimes with certain authorial modifications of classical approaches to sentiment analysis [14; 15].

From the point of view of machine learning, SA is a problem of binary classification into negative and positive classes (although sometimes there is a third class - neutral assessment [5]), so to automate it, you can use any classification methods, including logistic regression, Naïve Bayes, reference vectors and other techniques [16].

It is also essential to pay attention to the target of the study [17]. If the main task is to determine the subject of texts, rather than their tonalities, then the word bag approach remains effective, based on the frequency of specific tokens in the texts, and does not consider their order [18; 19]. Although the architecture of recurrent neural networks can account for the order, the relevant models do not understand the priority of different parts of the text relative to the target task. They cannot indicate that some words/sentences in the text carry essential information about sentiment and others do not. If the main task of the study requires taking into account the priority of words, it is solved by models-transformers, among which the BERT model used by the authors is the most popular, with such an essential element of architecture as attention-blocks [3]. These elements of the architecture of transformers are simple artificial neural networks. They perform the task of determining the importance of the token for the target function, which the authors successfully demonstrated on the example of analyzing the dynamics of tones of Ukrainian and Russian media reports about famous politicians. Some researchers have successfully combined conventional and computer content analysis, the so-called hybrid method [20].

In this paper, we focus on the latest type of neural network for word processing, because regardless of the approaches and methods used, researchers conclude that the neural network predicts sentiments most accurately [16] and demonstrate that the combination of neural network architectures is more effective for determining sentiments [21; 22].

The approach's effectiveness leaves open the question of search accuracy and greater compliance of its results with the user's intentions [23]. The accuracy of the problem of multiclassification of the text solves the combination of BERT model and Liblinear [24].

It is also worth emphasizing that the goal for sociology involves the construction of a conceptual model, research hypotheses, etc. Here, researchers appeal to the methodology of sociological research. The research methodology depends on its subject.

5. CONCLUSION

As part of the study, the authors demonstrated the applicability of new methods of sentimental analysis of texts to study the dynamics of the tones of Ukrainian and Russian media reports about famous politicians.

To this end, the authors identified a model of directed analysis of sentiments, the architecture of which involves working with local and global contexts, namely: the language model Bidirectional encoder for transformers (BERT). Taking into account the applied task, Ukrainian-language (youscan/ukr-roberta-base) and Russian-language (DeepPavlov/rubert-base-cased) models with F1-score quality indicators of 0.799 and 0.741 for each language, respectively, were trained.

Analysis of the dynamics of tones of messages of multilingual media using these models allowed to confirm the hypotheses about (1) greater criticism of the Ukrainian media to its power than the Russian media - to its own; (2) the dominance of negative tone in reports of foreign politicians in both media, with the tendency to increase the negative is characterized only by the UE; (3) the gradual loyalization of the Russian media to its power.

In general, the method of targeted sentiment analysis based on models-transformers showed the prospect of large-scale comparative studies of texts.

AUTHORS' CONTRIBUTIONS

S. Salnikova offered the logic of presenting this study, a conceptual model of the mediatization of politicians.

R. Kyrychenko performed machine analysis of texts. It was his responsibility to choose the BERT model, to train Ukrainian and Russian models.

ACKNOWLEDGMENTS

The authors are grateful to YouScan [8], whose development (a language model) we used to prepare this article.

REFERENCES

[1] Angelov, D. (2020), Top2Vec: Distributed Representations of Topics. *arXiv*, Homepage, available at: http://arxiv.org/abs/2008.09470

[2] Wang, J. Peng, B. & Zhang, X. (2018), "Using a Stacked Residual LSTM Model for Sentiment Intensity Prediction", *Neurocomputing*, vol. 322, pp. 93-101. DOI: 10.1016/j.neucom.2018.09.049

[3] Yang, X. Bian, J. Hogan, W. R. & Wu, Y. (2020), "Clinical concept extraction using transformers", *Journal of the American Medical Informatics Association*, vol. 27 (12), pp. 1935-1942. DOI: 10.1093/jamia/ocaa189

[4] Liu, B. (2012), Sentiment analysis and opinion mining, vol. 5 (1), pp. 1–167. DOI: 10.2200/s00416ed1v01y201204hlt016

[5] Nasukawa, T. & Yi, J. (2003), Sentiment analysis. DOI: 10.1145/945645.945658

[6] Strömback, J. & Esser, F. (2014), "Mediatization of politics: Towards a theoretical framework", in Strömback, J. & Esser, F. (Eds.), *Mediatization of politics: Understanding the transformation of Western democracies*, pp. 3–30, Basingstoke, UK: Palgrave MacMillan. DOI: 10.1057/9781137275844

[7] Zeng, B. Yang, H. Xu, R. Zhou, W. & Han, X. (2019), "LCF: A local context focus mechanism for aspect-based sentiment classification", *Appl. Sci.*, vol. 9 (16), 3389, DOI: 10.3390/APP9163389

[8] YouScan, Homepage, available at: https://youscan.io/blog/ukrainian-language-model/

[9] Kuratov, Y. & Arkhipov, M. (2019), Adaptation of deep bidirectional multilingual transformers for russian language, Homepage, available at: http://arxiv.org/abs/1905.07213

[10] Vos, D. & Van Aelst, P. (2018), "Does the Political System Determine Media Visibility of Politicians? A Comparative Analysis of Political Functions in the News in Sixteen Countries", *Political Communication*, vol. 35 (3), pp. 371-392. DOI: 10.1080/10584609.2017.1383953

[11] Zoizner, A. Sheafer, T. & Walgrave, S. (2017), "How Politicians' Attitudes and Goals Moderate Political Agenda Setting by the Media", *International Journal of Press-Politics*, vol. 22 (4), pp. 431-449. DOI: 10.1177/1940161217723149

[12] Wiedlack, K. (2020), "Enemy number one or gay clown? The Russian president, masculinity and populism in US media", *Norma*, vol. 15 (1), pp. 59-75. DOI: 10.1080/18902138.2019.1707459

[13] Rahman, A. Sadat, M. & Siddik, S. (2021), "Sentiment Analysis on Twitter Data: Comparative Study on Different Approaches", *International Journal of Intelligent Systems and Applications (IJISA)*, vol. 13(4), pp. 1–13. DOI: 10.5815/ijisa.2021.04.01

[14] Dadhich, A. & Thankachan, B. (2021), "Sentiment Analysis of AmazonPproduct Reviews Using Hybrid Rule-based Approach", *International Journal of Engineering and Manufacturing (IJEM)*, vol. 11 (2), pp. 40–52. DOI: 10.5815/ijem.2021.02.04

[15] Görmez, Y. Işık, Y. E. Temiz, M. & Aydın, Z. (2020), "FBSEM: A Novel Feature-Based Stacked Ensemble Method for Sentiment Analysis", *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 12 (6), pp. 11–22. DOI: 10.5815/ijitcs.2020.06.02

[16] Mostafa, G. Ahmed, I. & Junayed, M.S. (2021), "Investigation of Different Machine Learning Algorithms to Determine Human Sentiment Using Twitter Data", *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 13 (2), pp. 38–48. DOI: 10.5815/ijitcs.2021.02.04

[17] Kyrychenko, R. (2021), "Typology of Tasks of Machine Analysis of Texts in Contemporary Sociology", *Sociological Studios*, vol. 19(2). DOI: 10.29038/2306-3971-2021-02-41-48

[18] Rudkowsky, E. Haselmayer, M. Wastian, M. Jenny, M. Emrich, Š. & Sedlmair, M. (2018), "More than bags of words: Sentiment analysis with word embeddings", *Communication Methods and Measures*, vol. 12 (2-3): Computational Methods for Communication Science, pp. 140–157. DOI: 10.1080/19312458.2018.1455817

[19] Taboada, M. Brooke, J. Tofiloski, M. Voll, K. & Stede, M. (2011), "Lexicon-based methods for sentiment analysis", *Computational Linguistics*, vol. 37 (2), pp. 267–307. DOI: 10.1162/coli_a_00049

[20] Su, L.Y-F. Cacciatore, M.A. Liang, X. Brossard, D. Scheufele, D.A. & Xenos, M.A. (2017), "Analyzing public sentiments online: combining human- and computer-based content analysis", *Information Communication & Society*, vol. 20 (3), pp. 406-427. DOI: 10.1080/1369118X.2016.1182197

[21] Elzayady, H. Badran, Kh.M. & Salama, G.I. (2020), "Arabic Opinion Mining Using Combined CNN - LSTM Models", *International Journal of Intelligent Systems and Applications (IJISA)*, vol. 12 (4), pp. 25–36. DOI: 10.5815/ijisa.2020.04.03

[22] Ren, Z. Shen, Q. Diao, X. & Xu, H. (2021), "A sentiment-aware deep learning approach for personality detection from text", *Information Processing & Management*, vol. 58 (3), 102532. DOI: 10.1016/j.ipm.2021.102532

[23] Fu, Z. Wang, Y. Sun, X. Zhang, X. "Semantic and secure search over encrypted outsourcing cloud based on BERT", *Front. Comput. Sci.*, vol. 16 (2), 162802. DOI: 10.1007/s11704-021-0277-0

[24] Xue, Y. Wang, X. Gao, Z. (2019), "Multi-classfication Sentiment Analysis Based on the Fused Model", *IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 162802, pp. 1771-1775. DOI: 10.1109/ICTAI.2019.00265