Deception Detection in Online Media[⋆]

Alsu Zaynutdinova[1,2⋆⋆], Dina Pisarevskaya[3], Maxim Zubov[4⋆⋆⋆], and Ilya Makarov[1[0000−0002−3308−8825]]

[1] National Research University Higher School of Economics, Moscow, Russia
[2] Department of Data and Network Science, Central European University, Budapest
[3] FRC CSC RAS, Moscow, Russia
[4] National University of Science and Technology MISIS, Moscow, Russia
zaynutdinova_alsu@phd.ceu.edu,dinabpr@gmail.com,zubovmv@gmail.com,iamakarov@hse.ru

Abstract Russian Federation and European Union are fighting against fake news together with other countries in various topics. The disinformation affected British referendum of existing EU, the US election and Catalonia's referendum are broadly studied. A need for automated fact-checking increases, European Commission's Action Plan 8 is an evidence. In this work, we develop a model for detecting disinformation in Russian language in online media. We use reliable and unreliable sources to compare named entities and verbs extracted using DeepPavlov library. Our method shows four time greater recall compared to chosen baseline.

Keywords: Fake news · Information extraction · Fact checking · DeepPavlov · Named Entities

1 Introduction

Throughout the history of humanity, disinformation have been always with us. Since the Internet became our new tool of disseminating information, the spread of deception increased as well. It became a weapon for both economic and ideological reasons.

Deceptive stories flourished for two reasons: firstly, because the producers and curators of fake news content are able to monetise their content through advertising platforms from Facebook and Google; secondly, because social media platforms have effectively eliminated entry barriers to media production and distribution [1].

Although, certain deceit concept prevails in human social relations, it is always questionable whether lies can improve everybody's life. In some cases, false information may prevent conflict, but it will always be a method of manipulating decision making mechanisms. Knowing the fact of such manipulation goes against human nature and control system foundations.

[⋆⋆] Corresponding author

[⋆⋆⋆] The research was supported by the Russian Science Foundation grant 19-11-00281.

Nowadays false information shapes our reality: businesses collapse due to fake bad reviews, stock markets dramatically drop due to fake news or rumours related to companies [10], [6]. Due to lack of open information verification, people may buy defective products that can harm their lives. Moreover, people are being politically polarised [15], [16] and in this situation, disinformation can be spread by readers without reasonable judgement. It is even said that we are living in a post-truth era.

Most of works on automated deception detection have been made for English language [17]. As for Russian language, several papers have been written [14] [12]. In 2019, the Russian Federation adopted a bill that prohibits posting fake news in the media and the internet [4], nevertheless, it had not presented a methodology for detecting disinformation. We believe that some methodology of revealing fake news should be published so the bill would not function as a censorship. A comprehensive study on the concepts behind a lie can be found in [13].

Hence, our goal is to build a model to detect fake news in the Russian language. Our aim is to be able to detect prominent disinformation spread by specific websites.

By prominent disinformation we mean that it is relatively easy to detect for fact-checkers [8]. These type of disinformation usually appear in small websites, but can disseminate throughout Social Media. Usually, the bigger media organisations are not likely to write such news in order to preserve their reputation. However, by saying this, we do not presume that such media outlets would never publish disinformation. That being said, they would most definitely have more stringent fact checking processes in place, otherwise their reputation as a reliable source can be denigrated. Consequently, they would receive less financial support or advertisement moving forward. Although, we have to differentiate fake news from bias in media [2].

In our opinion, online media environment is very dynamic, therefore, using supervised machine learning models is not sufficient for our task. Also, using the knowledge that fact-checkers accumulated during their work is better for other tasks, like FullFact https://fullfact.org/ and Politifact https://www.politifact.com/ do. Prominent disinformation is written on actual topics and need to be debunked as soon as possible in order to prevent its spreading. Specifically, attempts for building argumentation systems can be also used for such systems in order to understand the fact-based systems reasoning [7,3].

2 Methodology

In this work, we want to predict whether news are real or false by comparing them to datasets of reliable and unreliable sources.

The reliability of source is based on two aspects. First, it is based on source's traffic. We suppose, that news outlets care about their reputation and they do not spread prominent disinformation. In other words, news organisations do not want to loose credibility in front of their audience. Therefore, we call them reliable sources in frames of this work. However, we do not claim that they will not write false news ever. We think, that in order to debunk disinformation spread by reliable sources, professional fact-checking is still required. It is a higher level of problem solving. Second, the reliability based on how many fakes were written by the news source. The more prominent fake news are posted, the more probability that the source would be called unreliable.

For selected reliable and unreliable sources, we focus on specific news outlets and wrote parser using BeautifulSoup library to retrieve data. The next step is to unify the structure of the data that had been retrieved from different sources. Then, we lemmatise news texts by Pymorphy2.

To preprocess the articles, we deleted the synopsis of the event. Usually this type of text is written at the end of a news article. For example, it can start with words like “Напомним” (We remind) , “Ранее сообщалось” (As reported earlier), etc.

We also manually created additional stop-words list for the news articles. This list contains, in particular, verbs, that often are used in news texts, such as “say”, “report”, “comment”, “reply” etc.

In the following steps we used DeepPavlov library [5]. We need to extract all named entities, verbs and numbers. To extract named entities, we used DeepPavlov’s component Named Entity Recognition. Then, we used Morphological Tagger to tag morphological entities like verbs and numbers. In the end, we had a list with lists that contains sets with retrieved words.

The idea of this methodology is that news of a similar nature should have the same named entities and at least 5% similar verbs and numbers (the threshold was chosen experimentally, but obviously, it should usually be non-zero).

3 Dataset

As we stated above, we use news from both reliable and unreliable sources. In order to hold neutrality and as to not create censorship, we took two media outlets regarded as media with different audiences. This is pivotal to our work, because the methodology can create censorship.

We collected news articles from news organisations like “Meduza” and “Russia Today” from 1st January to 16th April. “Meduza” is a non-governmental, private media outlet that is fully financed by advertisement. On the other hand, Russia Today is sponsored by the government. In the European Union’s Action Plan [5] against Disinformation, Russia Today is claimed to be a pro-Russian propaganda media source. But as our plan is to detect prominent fake news and not to create censorship, we consider Russia Today to be a reliable source. In total we have 3991 articles from “Russia Today” and 2400 from “Meduza”.

For deceptive news dataset, we used fact-checking “StopFake”. StopFake is a Ukrainian organisation that claims “Struggle against fake information about Ukraine”. The majority of fake news that they debunk is taken from Russian

[5] http://docs.deeppavlov.ai/en/master/intro/quick_start.html

media. However, some of scholars claim [9] that StopFake are biased in their approach. Taking this concern into consideration, we carefully selected 70 examples of disinformation that StopFake have debunked during the period from 1st January 2019 to 16th of April, 2019 (for which we see no controversial comments in other sources).

4 Baseline and Metrics

Firstly, we implemented a vectoriser of words TfidfVectorizer from Scikit-learn library for machine learning. This function transforms a matrix to a normalised TF-IDF representation of words. The reason why we did this was because we wanted to use classification models as a baseline. In order to use relatively simple classification models, we needed to convert our words frequencies into a Boolean matrix: ‘1’ means the word exist in a document and ‘0’ means it does not. Hence, we will have a matrix, in which in columns are documents and in rows all words used in the texts.

After representing our data into vectors, we can implement classification algorithms to create a baseline for our model, having 80%/20% train/test split. Cross-entropy loss function was weighted in order to balance classes.

We tried to use linear classification models like LogisticRegression, SGDClassifier, StratifiedKFold and Boost Gradient Machine. However, the results of the most of the models were not sufficient and provide low quality. The only classifier that we used that had some results is Boost Gradient Machine.

We used metrics like precision, recall and accuracy. In our opinion, for this task recall is more important than other metrics. We think so due to it is better to label a true news as false and then debunk it, rather than miss it and never be able to check it. The balance between precision and recall may be adopted to specific domain or purpose of use, however, the current results provide preliminary concept of applicability of suggested method, so we left this question for the future work.

5 Results

The recall of our model is 0.814, whereas baseline score is 0.185. We find this is promising result, thus we are planning to improve it in further studies.

As for the results, we obtained several findings. First, we built a dataset for truthful and false news. In total, we have 70 fake news and 6391 truthful news. Second, we built a baseline for our model, which is Boost Gradient Machine classifier. Finally, we have the results for our model presented in Table 1. The precision metrics are quite low, and high accuracy metrics does not provide meaningful results because of highly unbalanced classes in our case. Recall of our method was quite good, however it does not provide promising results due to lack of good data for the task and no semantic features were considered in the model.

Table 1. Results of experiments

Model	Recall	Precision	Accuracy
LogisticRegression	< 0.01	< 0.01	< 0.01
StratifiedKFold	< 0.01	< 0.01	< 0.01
SGDClassifier	< 0.01	< 0.01	< 0.01
XGBoost	0.18	0.22	0.98
Our method	0.81	0.1	0.72

We could retrieve the most frequent words that are used in false news. They are related to Russian-Ukrainian conflict.

The words are: [‘украина’, ‘переселенец’, ‘донбасс’, ‘лнр’ ,‘всу’, ‘марочко’, ‘украинский’, ‘язык’, ‘бригада’, ‘гривна’, ‘военнослужащий’, ‘милиция’, ‘крымчанин’, ‘подполковник’, ‘безвизовый’, ‘новотошковский’, ‘киев’, ‘керченский’, ‘учение’, ‘пункт’, ‘петра’, ‘батальон’, ‘штурмовой’, ‘эскалация’, ‘чёрный’, ‘мотороллер’, ‘акватория’, ‘море’, ‘ред’, ‘пролив’, ‘киевский’, ‘пьяный’, ‘горный’]

6 Limitations

First, this work is detecting prominent fake news. Therefore, if disinformation or misinformation were not prominent, the model would not find it. In that case manual fact-checking is required. Second, if the prominent news would be posted by a reliable source, the accuracy of the model will drop. Taking this into consideration, we will build a non-binary model in future.

The dataset of disinformation is really small. The reason for this is that the proportion of false news always be much smaller comparing to truthful news. If the deceptive information has been implemented within a text, for example, as a description of some detail, then it will not be selected as disinformation.

Although, TF-IDF applicability as a method is questionable, because it is invariant to semantic perturbations with “negative” statements and clausal sentence structure.As was mentioned by a reviewer, TF-IDF may work because usually fake news are not posted on “fictional” events with “negative” formulations. It is also quite easy to fool the algorithm based on TF-IDF if fake news publisher knows the principle behind the algorithm.

7 Future Work

Moving forward, we would strive to score sources dynamically. Initially, the credibility score would be given manually. Then, the sources would be compared by each other to calibrate the credibility score. The credibility percentage will vary on how many fake news were produced by a media organisation. This would be a dynamic and semi-automated process. We plan to improve our model and to add new features for extracting information from news article. We also want to look into whether someone has used satirical articles as a reliable source. There have been cases when articles from the satirical website Panorama [11] were used by some popular bloggers.

Acknowledgement

We thank all the reviewers of this work for their fruitful comments and discussion that we aim to take into account in our future studies on the topic.

References

1. Allcott, H., Gentzkow, M.: Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives **31**(2-Spring), 211–236 (2017)
2. Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., Nakov, P.: Predicting factuality of reporting and bias of news media sources. arXiv preprint arXiv:1810.01765 (2018)
3. Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., Panchenko, A.: Targer: Neural argument mining at your fingertips. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 195–200 (2019)
4. Duma: What is fake news and what is punishment for it?
5. European Commision: Action plan against disinformation (2018)
6. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Communications of the ACM **59**(7), 96–104 (2016)
7. Galitsky, B., Parnis, A.: Accessing validity of argumentation of agents of the internet of everything. In: Artificial Intelligence for the Internet of Everything, pp. 187–216. Elsevier (2019)
8. Hardalov, M., Koychev, I., Nakov, P.: In search of credible news. In: International Conference on Artificial Intelligence: Methodology, Systems, and Applications. pp. 172–180. Springer (2016)
9. Khaldarova, I., Pantti, M.: Fake news: The narrative battle over the ukrainian conflict. Journalism Practice **10**(7), 891–901 (2016). https://doi.org/https://doi.org/10.1080/17512786.2016.1163237
10. Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., et al.: The science of fake news. Science **359**(6380), 1094–1096 (2018)
11. Meduza: Сайт "Панорама" стал русским the onion
12. Pisarevskaya, D.: Deception detection in news reports in the russian language: Lexics and discourse. Proceedings of the 2017 EMNLP Workshop on Natural Language Processing meets Journalism p. 74–79 (2017)
13. Pisarevskaya, D., Galitsky, B., Taylor, J., Ozerov, A.: An anatomy of a lie. In: Companion Proceedings of The 2019 World Wide Web Conference. pp. 373–380. ACM (2019)
14. Pisarevskaya, D., Litvinova, T., Litvinova, O.: Deception detection for the russian language: Lexical and syntactic parameters. In: Proceedings of RANLP Natural Language Processing and Information Retrieval Workshop. p. 1–10. Varna, Bulgaria (2017), https://aclweb.org/anthology/papers/W/W17/W17-7701/
15. Spohr, D.: Fake news and ideological polarization: Filter bubbles and selective exposure on social media. Business Information Review **34**(3), 150–160 (2017)

16. Vicario, M.D., Quattrociocchi, W., Scala, A., Zollo, F.: Polarization and fake news: Early warning of potential misinformation targets. ACM Transactions on the Web (TWEB) **13**(2), 10 (2019)
17. Zhou, X., Jain, A., Phoha, V.V., Zafarani, R.: Fake news early detection: A theory-driven model. arXiv preprint arXiv:1904.11679 (2019)