Fake news detection for the Russian language

Gleb Kuzmin, Daniil Larionov,
Dina Pisarevskaya, Ivan Smirnov

The study was funded by Russian Foundation for Basic Research according to the research project No 17-29-07033.

Outline

I. Introduction
II. Related research
III. Datasets
IV. Experiments: pipeline and models: bag-of-n-grams, bag of Rhetorical Structure Theory features, and BERT embeddings
V. Results for best models
VI. Error analysis

Fake news for Russian

2020:

- The Russian legal system has special criminal law for spreading fake news about emergencies (http://duma.gov.ru/news/29982/, in Russian).
- Two fact-checking websites are created: https://provereno.media/ and https://fakecheck.ru/ . Blogging platform Yandex.Zen starts its fact-checking and misinformation detection program https://yandex.ru/support/zen/requirements/fact-checking.html
- Increasing number of fake news, connected to Russia

Our goals

- to compare different models for fake news detection in Russian: based on bag-of-n-grams, based on discourse features (Rhetorical Structure Theory (RST)), based on fine-tuned BERT for Russian;
- to check if satirical news should be singled out as a separate class, among fake news and real news, or can be combined with fake news.

Related research: linguistic features

- linguistic features for fake news detection were studied for English: n-grams, POS tags, readability and complexity features, psycholinguistic features from LIWC and other sources, syntax features, sentiment features etc. (Ajao et al., 2019; Baly et al., 2018; Karadzhov et al., 2017; Pérez-Rosas et al., 2018; Potthast et al., 2018; Rashkin et al., 2017 etc.). Discourse features (Rubin et al., 2015; Atanasova et al., 2019; Karimi and Tang, 2019);

- linguistic features for satire detection: POS features (Rubin et al., 2016; Yang et al., 2017; De Sarkar et al., 2018), psycholinguistic, readability and structural text features (Yang et al., 2017), sentiment scores and named entity features (De Sarkar et al., 2018), BERT (Levi et al., 2019).

Related research: fake news detection for Russian

- Few initial research studies on fake news detection, each one was based on a single dataset.
- Basic lexical, syntactic, and discourse parameters were examined in (Pisarevskaya, 2017). The impact of named entities, verbs, and numbers was investigated in (Zaynutdinova et al., 2019).

Available datasets for Russian: we took them all

1. Dataset from (Pisarevskaya, 2017): 174 texts, equal number of fake and truthful texts, parsed in 2015-2017 from Russian news sources (available upon request).
2. Dataset from (Zaynutdinova et al., 2019): 8867 texts, with 1366 fake and 7501 real news (available upon request).
3. Fake news dataset from the satire and fake news website https://panorama.pub/. This dataset is a part of the Taiga corpus for Russian, it is freely available at https://tatianashavrina.github.io/taiga_site/downloads. We have taken 1803 satirical texts.

Data description

We created 5 smaller datasets from the described data and used each of them for model training:
1. train and test parts - non-satirical fake news and real news (Fakes & Fakes) (9041 samples, test size is 20 % of the dataset);

2. train and test parts - satirical fake news and real news (Satira Fakes & Satira Fakes) (10136 samples, test size is 20 % of the dataset with fixed seed);

3. train part - satirical fake news and real news, test part - non-satirical fake news and real news (Satira Fakes & Fakes) (9476 samples, fixed test size with 174 samples);

4. train part - satirical and non-satirical fake news and real news, test part - non-satirical fake news and real news (Fakes + Satira Fakes & Fakes) (11676 samples, fixed test size with 174 samples);

5. train and test part - satirical and non-satirical fake news and real news, 3 class classification (Fakes + Satira Fakes & Fakes + Satira Fakes) (11676 samples, test size is 20 % of the dataset with fixed seed).

Baseline:

Bag-of-n-grams, with TF-IDF preprocessing, for the baseline models.

Preprocessing consists of removing control characters, removing http-like links, and optional lemmatization.

A subset of the most informative features was selected, before training the model (by computing ANOVA F-value for each feature).

Classification: Support Vector Machines with RBF kernel. Also, for a 3-class classification task (Fakes + Satira Fakes & Fakes + Satira Fakes) we trained a Logistic Regression based model for better model interpretability.

RST Features:

We used the automated discourse parser for Russian proposed in (Shelmanov et al., 2019).

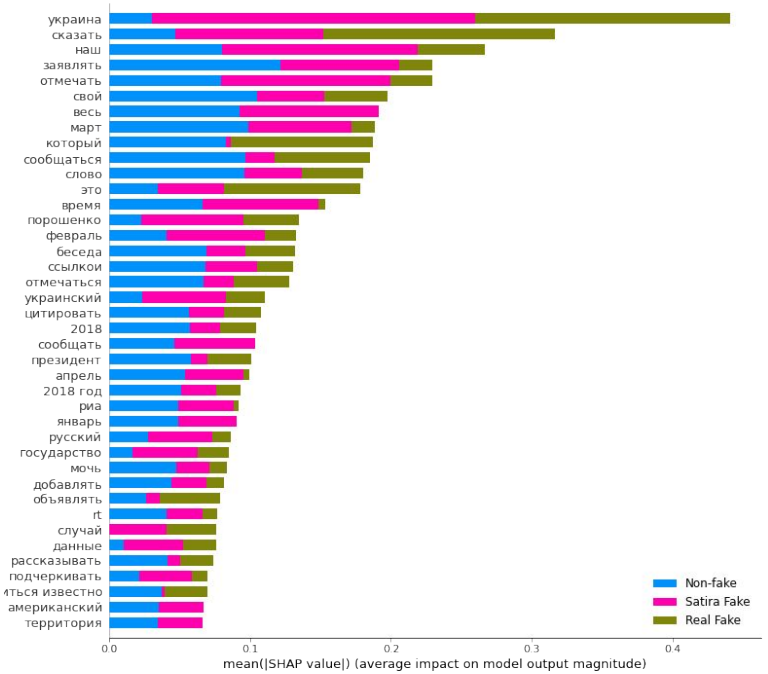
”Bag-of-rst” features: for each text, we have taken all the RST relations for all discourse units in the texts and encoded them into a one-hot vector. Such vectors were concatenated with feature vectors from the baseline model and used with an SVM-based classifier (Logistic Regression-based for the 3-class case).

Feature importance for bag-of-n-grams and ”bag-of-rst” models

We extracted feature importance using Shapley Additive explanations method.

Bag-of-n-grams:
the most important feature is the word ”Ukraine”.

This is because the Fakes part of our dataset is hugely based on Ukraine-related texts about the Russia-Ukraine conflict.

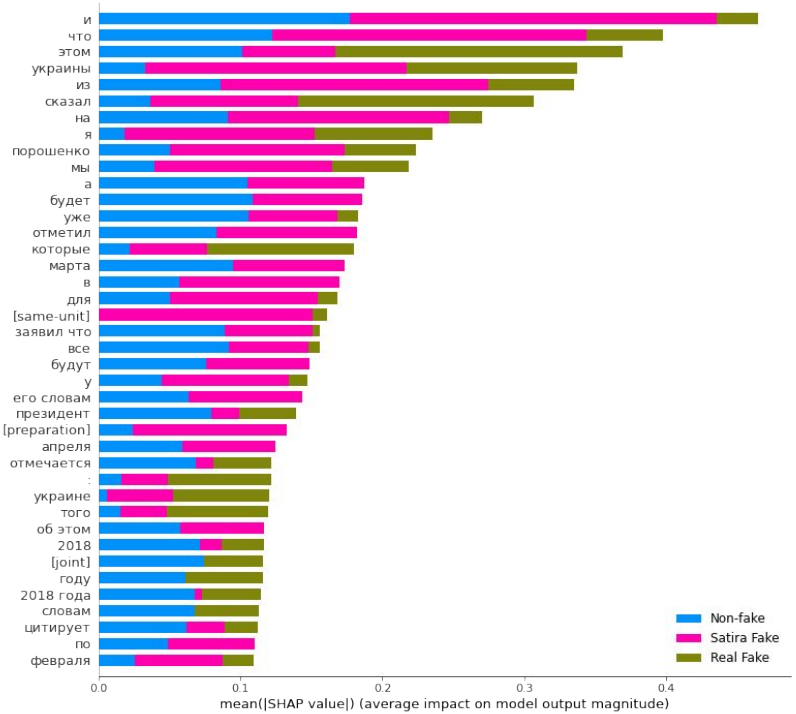


Feature importance for bag-of-n-grams and ”bag-of-rst” models

”Bag-of-rst”:

the presence of ”Same-Unit” or ”Preparation” relations almost always moves a model prediction towards ”Satire” class, while ”Joint” never does so.

All the top features on both model types are unigrams.



Fine-tuning BERT

We used pre-trained RuBERT for Russian from DeepPavlov (Burtsev et al., 2018) with Hugging Face.

In the process of fine-tuning, we trained only the last fully-connected layer with weighted cross-entropy as the loss function. That was done due to the unbalanced class distribution in our data.

As we used RuBERT for tuning, all our news texts were truncated at a size of 512 tokens. We found that 512 tokens are enough in the case of our dataset because only 1% of news has a length of more than 512 tokens.

RuBERT fine-tuning results

Dataset	F1-score, train/test	accuracy, train/test	roc-auc, train/test
1	0.765/0.778	0.883/0.890	0.745/0.752
2	0.881/0.887	0.906/0.909	0.891/0.895
3	0.446/0.333	0.806/0.500	0.500/0.500
4	0.715/0.546	0.738/0.546	0.718/0.546
5	0.741/0.748	0.823/0.822	0.913/0.909

Baseline (bag-of-n-grams) and RST features results

We achieved decent results both on binary classification datasets (1-4) and 3-class cases (5). RST features do not improve the performance of bag-of-n-grams models, although the RST-based model has RST features in the top-20 of the most important features.

Dataset	SVM-baseline	SVM "bag-of-rst"	LogReg-baseline	LogReg "bag-of-rst"
1	0.8800	0.8796	0.8875	0.8829
2	0.9576	0.9509	0.9513	0.9562
3	0.5950	0.5886	0.5919	0.5944
4	0.5600	0.5671	0.5576	0.5743
5	0.9084	0.8901	0.9076	0.9042

Binary classification or 3-class case?

Satire and fake news: together or not?

The performance on the dataset 4, where satire and real fakes are mixed, is worse than on the dataset 3, where the model is trained on satire texts and tested against real fakes. So we decided to separate satire texts, non-fake texts, and fakes into 3 different classes (dataset 5).

Comparison with human performance on the test set part

We already had a ground truth annotation for our datasets.

- But:

Additional manual annotation: about 500 random texts from the test set of our dataset for 3 class classification, in order to compare the results of our models with the human score on this part of the test set, and for checking the cases, where the models gave wrong predictions, more thoroughly.

Model	F1-score	accuracy
1 annotator	0.564	0.731
2 annotator	0.516	0.705
3 annotator	0.806	0.881
RuBERT	0.740	0.815
Best model (SVM)	0.908	0.941

Comparison with human performance and Error analysis

Metrics on the re-annotated test set part slightly differ from the metrics on the whole test set.
Results for the RuBERT model:

Class	Precision (whole set)	Precision (annotated)	Recall (whole set)	Recall (annotated)
Real news	0.867	0.849	0.895	0.908
Fake news	0.635	0.638	0.544	0.507
Satirical news	0.774	0.806	0.768	0.755

N-grams model metrics on the annotated test set part:

Class	Precision	Recall
Real news	0.932	0.956
Fake news	0.867	0.712
Satirical news	0.896	0.936

Comparison with human performance and Error analysis

’Gold’ labels and labels created by annotators: possible issues of concern.

- Only one annotator provided tags that were close to the ’gold’ labels (f1 score 0.806).
- Inter-annotator agreement between 3 annotators: substantial agreement only in distinguishing real news from fake and satirical news.
- It is more simple to detect satire. 71% satirical texts were annotated correctly, in comparison with 25% fake texts.
- Fake texts are mixed up with real news: in 2% cases, fake texts were labeled as satire by one single annotator, in other cases, they were labeled as fake or real ones.

Subjectivity of the manual approach to fake news detection.

Comparison with human performance and Error analysis

Inter-annotator agreement between 3 annotators: substantial agreement only in distinguishing real news from fake and satirical news.

Inter-annotator agreement for 498 texts (re-annotated part of the test set):

Agreement	Fleiss' kappa
3 classes: satirical, fake, real news	0.485
2 classes: 1) fake and real news 2) satirical news	0.553
2 classes: 1) real news 2) fake and satirical news	0.629

Error analysis: findings

It is hard to detect manually if a text is fake or real without additional information - facts and context that human annotators may be aware/not aware of.

Гендиректору Третьяковской галереи Зельфире Трегуловой объявлен выговор из-за похищения картины художника Архипа Куинджи. [...] Картина Архипа Куинджи «Ай-Петри. Крым» была похищена с выставки в Третьяковской галерее 27 января. Преступник снял полотно со стены на глазах у посетителей и беспрепятственно вынес его из здания. [...]

The General Director of the Tretyakov Gallery, Zelfira Tregulova, was reprimanded for the theft of a painting by the artist Arkhip Kuindzhi. [...] The painting by Arkhip Kuindzhi “Ai-Petri. Crimea” was stolen from an exhibition at the Tretyakov Gallery on January 27. The perpetrator took the canvas off the wall in front of the visitors and carried it out of the building without let or hindrance. [...]

Error analysis: findings

- Satirical texts can be detected manually better from real news without additional information, based only on their text: it contains absurd.

Единственный в России ми-го Яша сбежал из зоопарка в Перми после просмотра ток-шоу Андрея Малахова.«Яшу ещё в младенчестве нам доставила из Гималаев профессор Наталья Рокотова. Всё его семейство погибло из-за схода лавины. Он рос у нас и стал настоящим любимцем детворы», — рассказала смотритель зоопарка Ульяна Браун.По её словам, ми-го пристрастился к просмотру телевизора, стоявшего неподалеку от его клетки. Особенно нравились ему старые комедии и мультфильмы. [...] Профессор Мискатоникского университета, знаменитый миголог Алберт Уилмарт бросил все свои дела и вылетел в Пермь.

Yasha, the only one mi-go [a fictional race of extraterrestrials created by H. P. Lovecraft] in Russia, escaped from the zoo in Perm after watching a talk show by Andrey Malakhov. “Yasha was brought to us from the Himalayas by Professor Natalya Rokotova as a child. His entire family was killed by an avalanche. He grew up with us and became a real favorite of children”, said the zoo caretaker Ulyana Braun. According to her words, mi-go became addicted to watching TV, which was standing near his cage. He especially liked old comedies and cartoons. [...] Professor of the Miskatonic University [a fictional university from by H. P. Lovecraft books], the famous mi-gologist Albert Wilmart dropped all his business and flew to Perm.

Error analysis: findings

- Among 498 re-annotated texts, most texts about statistics and economics data or accidents are real. Only one such text is fake.

“Количество ДТП в Москве в 2018 году с участием такси: плюс 17% — количество раненых, плюс 25% — количество ДТП, плюс 14% — количество погибших”, — цитирует [...] РИА Новости.По словам Пронина, причинами такого роста могло стать в том числе увеличение числа заказов, продолжительности рабочего дня, слабый контроль за водительским составом. [...] В феврале телеканал «360» передавал, что в Московской области заявили о снижении смертности в результате дорожно-транспортных происшествий на 8,3% в 2018 году.

“The number of car accidents in Moscow in 2018 with the participation of taxis: plus 17% - the number of injured, plus 25% - the number of accidents, plus 14% - the number of deaths", according to RIA Novosti [...]. Pronin said that such issues as the rising number of orders, the length of the working day, and weak control over the drivers team could be the reasons for such an increase. [...] The TV channel "360" broadcasted In February that a decreased rate of deaths caused by car accidents (8,3% in 2018) was announced.

Error analysis: findings

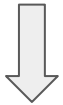
- Texts in all three classes can be biased: they may contain loaded language, opinion pieces, biased quotations. It might mislead the annotators.

«Мы абсолютно точно намерены довести экспорт иранской нефти до нуля», — уточнил он. Ранее Помпео сообщил, что Соединённые Штаты будут продолжать оказывать давление на Тегеран до тех пор, пока Иран не начнёт «вести себя как нормальная страна».

“We absolutely intend to bring the export of Iranian oil to zero”, he said. Earlier, Pompeo said that the United States will continue to exert pressure on Tehran until Iran starts “behaving like a normal country”.

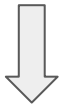
Error analysis: findings

- The datasets used in the study should be double-checked, to be unbiased. It concerns mostly texts with questionable quotations and texts with small fragments of fake content.



More proper annotation guidelines should be developed, i.e. to handle such cases: the quotation is correct, but it is not truthful.

- Among 498 annotated texts, there were no satirical texts about military news, so deceptive texts could be only fake.



The datasets should contain various topics and be taken from different sources, to avoid overfitting.

Conclusions

- The best BERT-based model achieved a 82.2% F1-score and 74.8% accuracy score on a 3 class classification task, which is bigger than the mean human result, but less than the metrics for the bag-of-n-grams based model, which achieved 90.8% F-score and 94.1% accuracy.
- The model outperforms human evaluation results based on the majority vote.
- ”Bag-of-rst” features do not improve the performance of the bag-of-n-grams model.
- Satirical news should be singled out as a separate class, among fake news and real news.

- Possible future work: claims verification module for Russian; collecting and annotating new social media datasets of fake, satirical, biased, and hyperpartisan news for Russian; multilingual sentence embeddings and transfer learning techniques.

Thank you for your attention!

Contact us please, if you have any questions or proposals:

Gleb Kuzmin (Moscow Institute of Physics and Technology kuzmin.gyu@phystech.edu)
Daniil Larionov (FRC CSC RAS dslarionov@isa.ru)
Dina Pisarevskaya (FRC CSC RAS dinabpr@gmail.com)
Ivan Smirnov (FRC CSC RAS ivs@isa.ru)

References

O. Ajao, D. Bhowmik, and S. Zargari. 2019. Sentiment aware fake news detection on online social networks. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2507–2511.

Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse information. ACM Journal of Data and Information Quality, 11(3):12:1–12:27.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3528–3539, Brussels, Belgium.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. In Proceedings of ACL 2018, System Demonstrations, pages 122–127, Melbourne, Australia.

Sohan De Sarkar, Fan Yang, and Arjun Mukherjee. 2018. Attending sentences to detect satirical fake news. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3371–3380, Santa Fe, New Mexico, USA.

References

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2017. We built a fake news / click bait filter: What happened next will blow your mind! In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pages 334–343, Varna, Bulgaria.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3432–3442, Minneapolis, Minnesota.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. Text-interdisciplinary Journal for the Study of Discourse, 8(3):243–281.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3391–3401, Santa Fe, New Mexico, USA.

Dina Pisarevskaya. 2017. Deception detection in news reports in the Russian language: Lexics and discourse.In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism, pages 74–79, Copenhagen, Denmark.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 231–240, Melbourne, Australia.

References

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2931–2937, Copenhagen, Denmark.

V.L. Rubin, N.J. Conroy, and Y.C. Chen. 2015. Towards News Verification: Deception Detection Methods for News Discourse. In Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, January 5-8.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In Proceedings of the Second Workshop on Computational Approaches to Deception Detection, pages 7–17, San Diego, California.

Artem Shelmanov, Dina Pisarevskaya, Elena Chistova, Svetlana Toldova, Maria Kobozeva, and Ivan Smirnov. 2019. Towards the data-driven system for rhetorical parsing of Russian texts. In Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019, pages 82–87, Minneapolis, MN, June.

Alsu Zaynutdinova, Dina Pisarevskaya, Maxim Zubov, and Ilya Makarov. 2019. Deception detection in online media. In Proceedings of the Fifth Workshop on Experimental Economics and Machine Learning at the National Research University Higher School of Economics co-located with the Seventh International Conference on Applied Research in Economics (iCare7), pages 121–127.