Fake news detection for the Russian language

Gleb Kuzmin*
Moscow Institute of Physics and Technology
/ Dolgoprudny, Russia
kuzmin.gyu@phystech.edu

Daniil Larionov*
Federal Research Center
"Computer Science and Control"
/ Moscow, Russia
dslarionov@isa.ru

Dina Pisarevskaya*
Federal Research Center
"Computer Science and Control"
/ Moscow, Russia
dinabpr@gmail.com

Ivan Smirnov
Federal Research Center
"Computer Science and Control"
/ Moscow, Russia
ivs@isa.ru

Abstract

In this paper, we trained and compared different models for fake news detection in Russian. For this task, we used such language features as bag-of-n-grams and bag of Rhetorical Structure Theory features, and BERT embeddings. We also compared the score of our models with the human score on this task and showed that our models deal with fake news detection better. We investigated the nature of fake news by dividing it into two non-overlapping classes: satire and fake news. As a result, we obtained the set of models for fake news detection; the best of these models achieved 0.889 F1-score on the test set for 2 classes and 0.9076 F1-score on 3 classes task.

1 Introduction

Fake news detection becomes a more significant task. It is connected with an increasing number of news in media and social media. To prevent rumors and misinformation from spreading in this news, we need to have a system able to detect fake news. It also could be useful because it's hard for people to recognize fake news. Approaches to this task are being developed for English. However, fake news texts can be written originally in different 'source' languages, i.e. in Russian. To tackle such content in social media, multilingual systems might be used. For Russian, only preliminary research for automated fake news detection was undertaken before. Moreover, fake news detection in Russian becomes more actual due to the appearance of new laws in the Russian legal system. For example, since 2020 the Russian legal system has special criminal law[1] for spreading fake news about emergencies. Furthermore, due to an increasing number of fake news, connected with Russia, it will be useful to have the fake news detection module for the Russian language to check original news in Russian.

In our paper, we trained and compared different models for fake news detection in Russian. We checked if various language features alone can be helpful for this task. As baseline models, we used Support Vector Machines (SVM) and Logistic Regression over a bag-of-n-grams. Also, we trained similar models, but over features obtained from the discourse parsing of news (Rhetorical Structure Theory (RST) discourse features, as in (Mann and Thompson, 1988)). For each news text, these features were constructed from its hierarchical discourse tree representation. It contains discourse (rhetorical) relations between text segments – discourse units, starting with the smallest 'leaf' segments - elementary discourse units. As the third model, we fine-tuned BERT (for the Russian language) for the fake news detection task.

* - These authors contributed equally.

[1] http://duma.gov.ru/news/29982/ (in Russian)

Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM), pages 45–57
Barcelona, Spain (Online), December 12, 2020.

We investigated all existing datasets for automated fake news detection for Russian. During our study, we also examined some hypotheses about different types of fake news. We divided our data into three parts - non-satirical fake news (that are equal to simple fake news), satirical fake news, and real news. The satirical fake news is the deceptive news, that was written deliberately with humorous purpose or at least without disinformation purpose. It is presented in the format and style of legitimate news articles (Rubin et al., 2016; De Sarkar et al., 2018). A more detailed description of the data splitting can be found in section 3. We established that, for the classification task, satirical news should be singled out as a separate class, among fake news and real news. Finally, we annotated the part of the dataset manually and compared the results of our models with human performance, setting goals for future research steps.

2 Related Work

1. Linguistic features. For English, linguistic features for fake news detection were studied during recent years. Content features are used in (Rashkin et al., 2017; Volkova et al., 2017; Ruchansky et al., 2017; Ma et al., 2018; Kochkina et al., 2018; Khattar et al., 2019). Text-based only approach is also used in (Ajao et al., 2019) (sentiment features), (Dungs et al., 2018; Wu et al., 2019) (stance detection). (Baly et al., 2018) suggest linguistic features (POS tags, sentiment scores, readability and subjectivity features, number of cognitive process words etc.) as well as source credibility features. (Karadzhov et al., 2017; Pérez-Rosas et al., 2018; Potthast et al., 2018) explore linguistic features, including ngrams, POS tags, readability and complexity features; psycholinguistic features from LIWC (Rashkin et al., 2017; Pérez-Rosas et al., 2018) and other sources (Rashkin et al., 2017), syntax features (Pérez-Rosas et al., 2018). There are several datasets, i.e. the dataset proposed in (Rashkin et al., 2017).

As for discourse features, (Karimi and Tang, 2019) incorporate hierarchical discourse-level structures for fake news detection. Structure-related properties (number of leaf nodes, preorder difference, parent-child distance) identify structural differences between fake and real news texts. The approach, based on discourse dependency trees, yields better results (82.19%) than approaches based on n-grams, on LIWC features, on RST discourse features taking the hierarchical structure of document into account (as in (Rubin and Lukoianova, 2015)), or BiGRNN-CNN and LSTM approaches using word embeddings in sentences. (Atanasova et al., 2019) propose a set of various features for context and discourse modeling. There are also discourse features among them, based on automated discourse parsing according to RST. They are focused on the direct relationship between a target sentence and other sentences in a segment and on the internal structure of a target sentence (number of nuclei and satellites). Such rhetorical relation types as Background, Enablement, Elaboration, Attribution are associated with factually-true examples.

Linguistic features for fake news detection have limitations and can be used in addition to automated fact-checking. I.e., Shuster (Schuster et al., 2020) study linguistic (stylistic) features for machine-generated misinformation detection and conclude that they are limited in detecting if texts generated by language models are fake, as such texts hide stylistic differences between falsified and truthful content.

Automated satire and biased text detection are closely connected tasks. For satire detection, absurdity feature (unexpected introduction of new named entities within the final sentence) (Rubin et al., 2016), 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) can be used. Satire can be distinguished from fake news using semantic representation with the BERT language model and with linguistic features based on textual coherence metrics that also include basic language features, such as readability features, sentence length, number of words (Levi et al., 2019). Language features, without checking external sources of information, are also mainly used for biased language and propaganda detection (Potthast et al., 2018; Da San Martino et al., 2019)

2. Claims verification. In automated fact-checking for English, (Thorne et al., 2018; Nie et al., 2019; Augenstein et al., 2019; Zhong et al., 2020; Portelli et al., 2020; Kochkina and Liakata, 2020) consider evidence detection and claims verification. Several recent studies are focused on evidence detection for explainable claim verification, based on semantic entailments for claims (Hanselowski et al., 2018; Ma et al., 2019), incorporating semantic similarity between comments and claims (Wu et al., 2020). BERT language model (Devlin et al., 2019) can be used for checking if a comment is

factual (Stammbach et al., 2019), for retrieving evidence sentences and verifying the truthfulness of claims against the retrieved evidence sentences (Soleimani et al., 2020); there are also experiments on using it solely, without any external knowledge or explicit retrieval components (Lee et al., 2020). There are several datasets, i.e. PHEME (Zubiaga et al., 2016), LIAR (Wang, 2017), RumourEval (Derczynski et al., 2017), FEVER (Thorne et al., 2018).

For Russian, few initial research studies on fake news detection have been conducted, each one 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). There are only a few datasets for fake news language and no existing datasets for claims verification. To build a misinformation detection system for social media posts in Russian, firstly we study language features for fake news detection. As discourse features were already regarded for fake news detection for Russian, we check their impact in our study too.

3 Data

3.1 Data Collection

To the best of our knowledge, there are three available datasets for fake news detection in Russian.

- 174 texts from (Pisarevskaya, 2017), with an equal number of fake and truthful texts, parsed in 2015-2017 from Russian news sources. The dataset is available upon request.

- texts from (Zaynutdinova et al., 2019). Total 8867 texts, with 1366 fakes and 7501 real ones. The dataset is also available upon request.

- 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.

We incorporate them for our research and base it on them. In our dataset, we used only one source of satirical news because we found only one reliable source of satirical fake news. It could induce some bias during model training. On the other hand, news from this source are written by different authors on various topics, so we suppose that one source could be enough.

3.2 Data Description

We created 5 smaller datasets from the described data and used each of them for model training. They are structured as follows:

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).

In datasets 1-4 we were solving a binary classification problem, therefore, satirical and non-satirical fake news was treated as one class. In the 5th dataset, we considered the multiclass classification, therefore satirical and non-satirical fakes were treated as two different classes. We specially created 5 datasets to check some hypotheses on fake news structure. First of all, we would like to test if satirical fake news is different from non-satirical fake news. To test this hypothesis, we used datasets 2-5, while the first dataset contains only non-satirical fake news and real news and serves as a reference score for comparison.

4 Experiments

4.1 Baseline

We chose a basic method of a 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. Also, we chose to select a subset of the most informative features before training the model. This was done by computing ANOVA F-value for each feature and selecting **k** highest scored features, where **k** is a hyperparameter.

Within this framework, we trained a classification model, based on 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. We employed Bayesian hyperparameter optimization (Snoek et al., 2012) with Hyperband (Li et al., 2016) early termination algorithm, in order to estimate optimal hyperparameter values (Table 8).

4.2 RST Features

For more advanced models, we employed features obtained from the RST parsing of our texts. We used the automated discourse parser for Russian firstly proposed in (Shelmanov et al., 2019). The first approach is a so-called "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). We employed the aforementioned hyperparameter optimization algorithm for this pipeline as well.

The second approach, aimed to better consider the hierarchical structure, was based on averaging embeddings of different discourse units. Firstly, we grouped nodes of the discourse tree by their relation. Then texts from both leaves of each node were concatenated and passed through the BERT-based sentence embedding model from (Kuratov and Arkhipov, 2019). Resulting vectors were averaged for each type of RST relation (see Table 1). We used the concatenation of averaged vectors together with the SVM based classifier.

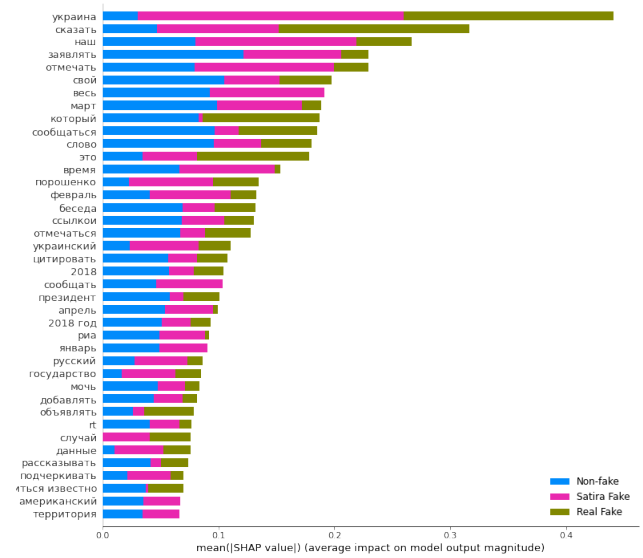
Relation		
attribution	background	cause-effect
concession	condition	contrast
elaboration	interpretation-evaluation	joint
preparation	purpose	same-unit
solutionhood		

Table 1: List of RST relations extracted by (Shelmanov et al., 2019)

However, the model from the second approach was not able to learn to predict anything from this set of features. It looks like the model always chooses to predict a single class i.e. works as a constant predictor. Thus, we chose to focus on bag-of-rst and BERT based models.

4.3 Feature Importance

For baseline and "bag-of-rst" models we extracted feature importance using Shapley Additive explanations method from (Ribeiro et al., 2016). We can see on charts (Figure 1) that the most important feature



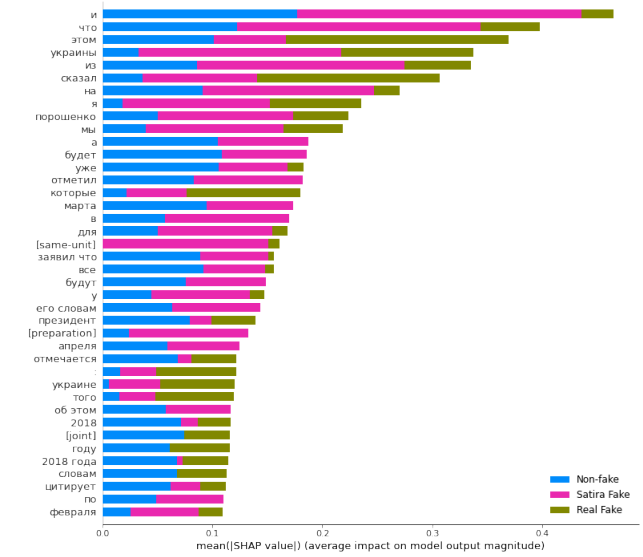


Figure 1: The most important features for the Logistic Regression-based baseline model and the ”bag-of-rst” model

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. Thus, this is a rich area for fakes generation by both sides of the conflict. Also, another Ukraine-related feature in the list is ”Poroshenko”, the surname of the former Ukrainian president Petro Poroshenko. Another conclusion that can be drawn from these features is that RST relations are among the most important features for the ”bag-of-rst” classifier. And these relations impacts all classes, the classifier does not overfit on them but uses these features together with bag-of-n-grams features. Speaking on RST relations, we see several examples: the presence of ”same-unit” or ”preparation” relations almost always moves a model prediction towards ”satire” class, while ”joint” never does so. ”Same-unit” is less informative as an utility relation (Shelmanov et al., 2019).

4.4 Ensembles

To further improve the quality of our models, we decided to the evaluate ensembles of best models. Firstly, we chose to implement an ensemble of baseline and 'bag-of-rst' models. The weight coefficients were additionally optimized for maximum performance. However, the ensemble only slightly outperforms the 'bag-of-rst' model by 0.001-0.002 of F1-score. At the same time, the ensemble of BERT models shows F1-score 0.02 less than just the best BERT model.

4.5 Fine-tuning BERT

In the last few years, BERT-based models showed state-of-the-art results in various NLP tasks. Particularly, these models also showed good results in sequence classification tasks. As one of the models, we used pre-trained RuBERT[2] from DeepPavlov (Burtsev et al., 2018) with Hugging Face (Wolf et al., 2019). 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. For tuning, we also used Adam optimizer with the linear scheduler (decays the learning rate of optimizer by γ each γ-steps). As we used BERT for tuning, all our news texts were truncated at a size of 512 tokens. Nevertheless, 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. More detailed model parameters are described in the results section alongside its performance.

5 Results

5.1 BERT

For the model trained on datasets 1-4 we used the following model parameters: batch size - 8, epochs - 20, lr - $9.2 \cdot 10^{-5}$, max_tokens - 512, γ - 0.357, γ-steps - 9. For 3 classes classification on dataset 5 we used other parameters: batch size - 8, epochs - 20, lr - $9.98 \cdot 10^{-5}$, max_tokens - 512, γ - 0.436, γ-steps - 8 (Table 2).

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

Table 2: RuBERT fine-tuning results

5.2 Baseline and RST Features Results

After stages of bayesian optimization, we found the final sets of hyperparameters (Table 9).

We achieved decent results both on binary classification datasets (1-4) and 3-class cases (5). As we can see in Table 3, RST features do not improve the performance of bag-of-n-grams models. However, as we see in Figure 1, the RST-based model has RST features in the top-20 of the most important features, thus such a model learns differently and uses the discourse structure for text scoring.

It is worth mentioning that we discovered strong negative correlation between the f1-score of the model and the minimal n-gram size, used for tokenization. That's why we always have an n-gram range of (1, ..). Also, all the top features on all the models are unigrams (see Figure 1). That's why we state that fake texts may be characterized by a very specific vocabulary and discourse structure, used by their authors.

[2] https://huggingface.co/DeepPavlov/rubert-base-cased-sentence

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

Table 3: F1-scores for baseline and RST feature-based models results for binary tasks and 3-class case

5.3 Difference between Satirical Fakes and Fakes

Our initial hypothesis was that satire is similar to fake news, thus it can be used to improve classification performance. Indeed, satire texts are much easier to obtain, because they are typically hosted on well-known satire-news media. And people are equivalently likely to misinterpret satire texts as truthful ones. However, our experimental evidence shows that satire significantly differs from real fakes. According to Table 2 and Table 3, the performance on the 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. Thus, we decided to separate satire texts, non-fake texts, and Fakes into 3 different classes (dataset 5).

5.4 Comparison with Human Performance

In this section, we compared the performance of our models with the human score. Although the datasets were already annotated before publishing them, for this task we manually labeled, in addition, about 500 random texts from the test part of our dataset for 3 class classification, to investigate the results in detail. Texts were labeled by 3 annotators, the guidelines were short and not explicit: we aimed to check how human annotators, not experts, define if news texts are fake, satirical, or real. It is worth noting that we already had a ground truth annotation for our datasets. So this additional manual annotation was only used for comparison of our models with the human score on the part of the test set, and for checking the cases, where the models gave wrong predictions, more thoroughly.

Then we used this manually labeled data as an additional (separate) test set for our models. The results are shown in Table 4.

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

Table 4: Comparison of our models with human performance

6 Error Analysis and Discussion

As an automated fake news detection system could be used as a preliminary filter before human evaluation, it is important to reduce the number of false positives for fakes and satire. For the selected BERT model, the false positives rate is only 0.05 for fakes and 0.06 for satire on the test set (0.05 and 0.05 on its human-annotated part). For the n-grams model, the false positives rate is 0.01 for fakes and 0.02 for satire on the test set (0.02 and 0.03 on its human-annotated part).

We performed an error analysis on the annotated part of the test set, to compare the model's results with human assessments. Metrics on this part slightly differ from the metrics on the whole test set (Table 5). Human annotators were also unsure of labeling satirical texts that were mispredicted as fake news. Real texts, that were mispredicted by the model as fake news, were mostly labeled correctly by the annotators but contained loaded language, emotional lexicon. 70% of them were about the politics of Ukraine (see

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

Table 5: BERT metrics on the test set

Section 4.3), yielding that dataset topics should have been more diverse. Real texts mispredicted as satire were about unusual events and required some additional knowledge, all of them were predicted correctly by the annotators with 100% inter-annotator consistency to satirical texts and, to a greater degree, fake texts that were mispredicted as real news, in some cases, it is hard to understand why they are not real without real-world knowledge, and annotators provide different labels. In these cases, people also can be mistaken.

We also looked at human labels and n-grams model results on the annotated part of the test set (Table 6). In cases where the n-grams model mispredicted fake news as satirical news and satirical news as fake news, human annotators also did not reach agreement: the texts were about politics and required the knowledge of facts. For all fake texts that were labeled as real by the model, the majority vote of annotators would also provide the 'fake' tag.

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

Table 6: N-grams model metrics on the annotated test set part

We investigated 'gold' labels and labels created by annotators and figured out possible issues of concern.

Only one annotator provided tags that were close to the 'gold' labels (f1 score 0.806). While checking the inter-annotator agreement between 3 annotators (about 500 texts), we found out that the annotators reached a substantial agreement only in distinguishing real news from fake and satirical news (Table 7).

According to the majority vote, 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. It also yields the subjectivity of a manual approach to fake news detection.

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

Table 7: Inter-annotator agreement for 498 texts

We also examined that:

1. 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 (examples: news about domestic policy in different countries). Therefore annotations might become subjective and it is harder to estimate the model's results. The claims verification module for Russian should be the next step of the research;

2. satirical texts can be detected manually better from real news without additional information, based

only on their text: it contains absurd (examples: news about an alien in the Zoo of Perm city (Russia), news about a bank payment terminal on the International Space Station);

3. among 498 annotated texts, most texts about statistics and economics data or accidents are real (examples: news about currency exchange markets and oil prices, news about car accidents statistics). Only one such text is fake. This peculiarity may differ in the real-world data;

4. texts in three classes can be biased: they may contain loaded language, opinion pieces, biased quotations (examples: news about politics in the Middle East, news about Russia-Ukraine conflict). Further research might be focused on hyperpartisan and satirical news and biased language detection;

5. 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;

6. 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.

7 Conclusions

The research can be considered as the first step in building an automated state-of-the-art system for Russian that could detect fake and satirical news and perform automated fact-checking. In this step, we studied language features for fake news and satire detection. We trained and compared different models for fake news detection in Russian, based on all existing available datasets for the Russian language. We investigated bag-of-n-grams features, bag of RST features, and BERT embeddings. We found out that satirical news should be singled out as a separate class, among fake news and real news. We also compared the score of our models with the human score.

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. We found that "bag-of-rst" features do not improve the performance of the bag-of-n-grams model, as they have reached almost the same scores on the test set. Further feature importance analysis and hyperparameter results analysis showed that unigrams are the most important features for fake news detection on our dataset. The model outperforms human evaluation results based on the majority vote.

For automated fake news detection, a combination of different methods should be applied. In future studies, the claims verification module for Russian should be developed and used together with the linguistic features models. New social media datasets of fake, satirical, biased, and hyperpartisan news for Russian should be collected and annotated according to detailed guidelines. We are also planning to try self-supervision methods for extending the datasets. After that, experiments should be performed on creating models for biased and hyperpartisan content, satirical content detection. Wider sets of language and content features should be used for them. We are also going to use multilingual sentence embeddings and transfer learning techniques, in order to incorporate the existing models and approaches, developed for automated claims verification for English, to this task for Russian.

Acknowledgements

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

The research was carried out using infrastructure of shared research facilities CKP "Computer science" of FRCCSC RAS. Regulations of CKP "Computer science" // Available at: http://www.frccsc.ru/ckp (date of the application 01.11.2020).

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.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. 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)*, pages 4685–4697, Hong Kong, China.

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.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. 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)*, pages 5636–5646, Hong Kong, China.

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.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370, Santa Fe, New Mexico, USA.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium.

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.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *WWW '19: The World Wide Web Conference*, pages 2915–2921, San Francisco, CA, USA.

Elena Kochkina and Maria Liakata. 2020. Estimating predictive uncertainty for rumour verification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6964–6981, Online.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA, August.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Nayeon Lee, Belinda Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41, Online.

Or Levi, Pedram Hosseini, Mona Diab, and David Broniatowski. 2019. Identifying nuances in fake news vs. satire: Using semantic and linguistic cues. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 31–35, Hong Kong, China.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. 2016. Hyperband: A novel bandit-based approach to hyperparameter optimization.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2561–2571, Florence, Italy.

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.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6859–6866.

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.

Beatrice Portelli, Jason Zhao, Tal Schuster, Giuseppe Serra, and Enrico Santus. 2020. Distilling the evidence to augment fact verification models. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 47–51, Online.

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.

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.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Victoria L. Rubin and Tatiana Lukoianova. 2015. Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology*, 66(5):905–917.

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.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *CIKM '17: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806.

Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Comput. Linguistics*, 46(2):499–510.

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.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol 12036*, pages 359–366.

Dominik Stammbach, Stalin Varanasi, and Guenter Neumann. 2019. DOMLIN at SemEval-2019 task 8: Automated fact checking exploiting ratings in community question answering forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1149–1154, Minneapolis, Minnesota, USA.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Lianwei Wu, Yuan Rao, Haolin Jin, Ambreen Nazir, and Ling Sun. 2019. Different absorption from the same sharing: Sifted multi-task learning for fake news detection. 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)*, pages 4644–4653, Hong Kong, China.

Lianwei Wu, Yuan Rao, yongqiang zhao, Hao Liang, and Ambreen Nazir. 2020. DTCA: Decision tree-based co-attention networks for explainable claim verification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1024–1035, Online.

Fan Yang, Arjun Mukherjee, and Eduard Dragut. 2017. Satirical news detection and analysis using attention mechanism and linguistic features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1979–1989, Copenhagen, Denmark.

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.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE11(3)*, pages 1–29.

Appendix A. List of hyperparameters.

Hyperparameter	Distribution/Range
C	uniform(0.001, 10)
gamma	uniform(0.0001, 1)
minimal n-gram frequency	uniform(1, 10)
n-gram's range	uniform, from 1..5-gram to 1..20-gram)
tokenization	with or without lemmatization
k	uniform(100, 10000)
max number of iterations	uniform(100, 10000)
penalty(for logreg)	l1 and elasticnet

Table 8: Hyperparameter space for SVM based classifier with RBF kernel and Logistic Regression based classifier

Model	N-gram range	topK	min frequency	C	γ	penalty
SVM - binary - baseline	(1, 3)	1048	1	2.871	0.8217	-
SVM - binary - RST	(1, 2)	5830	1	9.294	0.321	-
LogReg - binary - baseline	(1, 2)	2640	1	9.975	-	l1
LogReg - binary - RST	(1, 2)	2760	1	9.415	-	elasticnet
SVM - 3 class - baseline	(1,2)	9745	10	9.126	0.4708	-
SVM - 3 class - rst	(1, 15)	8194	8	9.940	0.1729	-
LogReg - 3 class - baseline	(1, 5)	8671	8	9.681	-	elasticnet
LogReg - 3 class - rst	(1, 4)	8411	9	9.781	-	l1

Table 9: Hyperparameter values