Fake News Detection
Project Proposal for NLP Course, Winter 2022

Marcel Affi
marcel.affi98@gmail.com

Mateusz Wójcik
mateusz.wojcik5.stud@pw.edu.pl

Arkadiusz Zalas
arkadiusz.zalas.stud@pw.edu.pl

Abstract

In this report we describe our initial research towards solving a problem of fake news detection which nowadays is a large issue. We explore already existing solutions for misinformation related problems and present datasets we will work with. Our aim is to compare different approaches by expanding basic preprocessing methods explicitly for entire article fake news detection and testing a few selected architectures.

1 Introduction

This project aims to detect news that contain misinformation, particularly fake news, based on e.g. online article texts. Since nowadays anyone can run a blog and write whatever they please, it is crucial to be able to differentiate between so-called real and fake news. We will test and compare different methods for misinformation detection (MID) using multiple datasets.

2 Background and Goals

Before introducing the existing research and the proposed methodology, let us start with discussing the formal definition of misinformation and its types.

2.1 What Is Misinformation?

Misinformation is a false statement or set of statements which mislead other people by hiding or "twisting" the actual facts (Islam et al., 2020). This concept causes feelings of mistrust, which is a growing problem in society due to increased popularity of the Internet – social media, news web sites, or gossip portals. It is usually a difficult task to distinguish between false and true statements, in which the mislead is very subtle. Mathematically, misinformation may be defined as a binary function, which assigns *true* or *false* to an arbitrary sentence. Nonetheless, the truth may be somewhere in between.

$$M(a) = \begin{cases} 1, & \text{if } a \text{ is true} \\ 0, & \text{if } a \text{ is false} \end{cases}$$

There are usually five different types of information specified:

- false information – broad concept of misinformation. It is intentionally used to be defined as a correct information interchangeably
- rumor – a story of doubtful truth, which is usually spread widely
- **fake news** – modified version of an original news or piece of information that is spread intentionally and usually very difficult to identify
- spam – an unwanted message with irrelevant, inappropriate, or even harmful information used to mislead users
- disinformation – false facts that are conceived to deceive a user

Type	Characteristics	Objectiveness	Severity	Integrity
Rumors	Ambiguous	Not sure	Low	Not sure
False information	Deception	Yes	High	False
Fake news	Misguided	Yes	Medium	False
Spam	Confused	Yes	Low	Not sure
Disinformation	Mislead/deceive	Yes	Medium	False

Figure 1: Different misinformation types considered in (Islam et al., 2020)

2.2 Impact of Misinformation

The phenomenon of all types of misinformation may influence various spheres of real world, for example society, economy, politics, emergency response during natural disasters, epidemic, and crisis. Public opinion is misled to confuse society and influence decisions and choices. Nowadays, due to the popularity of social network platforms like Facebook, Twitter, or Reddit, it has become easier to spread misinformation in seconds.

2.3 Our Goals

Our goal is to find potential patterns in fake news and analyse their structure so that they may be detected more accurately. We want to implement and test different approaches which originate from the existing research to propose new preprocessing pipelines or original improvements to known architectures.

3 Literature review

There were many approaches to tackle the problem of fake news detection. In (Islam et al., 2020) three different categories of models were described: discriminative, generative and hybrid models.

3.1 Discriminative models

There are three main discriminative models generally used for text classification: Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Recursive Neural Networks (RvNN). CNN has been successfully used in various text analyses (Jacovi et al., 2018). For the similar task of stance detection and rumor verification, word matrices consisting of word vectors were used to feed convolutional networks (Chen et al., 2017b). One research extended the usage of CNN to include not only text but also visual information such as images. This was accomplished by running two parallel CNNs for latent feature extraction, as fake news often also includes a wide variety of deceptive images (Yang et al., 2018).

RNNs are also used due to the sequential characteristics of text data. One of the approaches utilizes a combination of word2vec model architecture (Mikolov et al., 2013) with LSTM model to both encode and process rumor sentences for classification task (Alkhodair et al., 2020). Further extension use attention-based mechanisms that search for relevant parts of source sentence for predicting target label (Chen et al., 2017a).

The last considered model architecture is RvNN, which is used to capture the hierarchical structure of the input. One of the research describes the successful usage of the model for rumor detection on the Twitter dataset, which takes into consideration replies from other users' reactions, such as approving or disagreeing with the original post(Ma et al., 2018) (Zubiaga et al., 2016).

3.2 Generative models

The nature of generative models is well suited for a topic such as misinformation, as authors tend to intentionally sound as factual as possible to deceive the reader, the former playing the role of a generator and the latter indirectly acting as a discriminator. That said, there are many types of generative models that are applicable for this type of problem, most notably Restricted Boltzmann Machine, Deep Belief Network, Deep Boltzmann Machine, Generative Adverserial Network and Variational Autoencoder.

Starting with Deep Belief Networks which is a generative graphical model composed of multiple layers of latent variables, these types of networks can be interpreted as a composition of simple, unsupervised networks where each subnetwork's hidden layer serves as the visible layer for the next. It has been proved by Wei et al. (Wei et al., 2018) that this model achieves a better result than the traditional SVM-based approach.

Moving on to Generative Adverserial Networks, we observe that widespread rumors and misinformation usually result from the deliberate dissemination of information which is generally aimed at forming a consensus on rumor news events, Ma et al. (Ma et al., 2019) proposed a generative adversarial network model to make automated rumor detection more robust and efficient and is designed to identify powerful features related to uncertain or conflicting voice production and rumors.

Finally, we consider Variational Autoencoder models which make strong assumptions concerning the distribution of latent variables, they introduce a specific estimator for the training algorithm known as the stochastic gradient variational bayes (SGVB) estimator. Qian et al. (Qian et al., 2018) proposed this model to extract new patterns by analyzing a user's past meaningful responses on true

and false news articles, this played a vital role in detecting misinformation on social media. Additionally Wu et al. (Wu et al., 2017) explored whether the knowledge from the historical data analysis can benefit rumor detection. The result of their study was that similar rumors always produce the same behaviors.

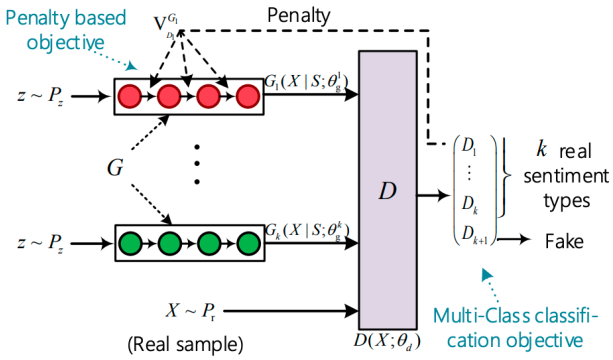

Figure 2: SentiGan architecture

We chose to implement the (Ke Wang, 2018) architecture proposed by Ke Wang and Xiaojun Wan as the model is focused more on the multi class classification objective, this will be useful especially for out LIAR dataset which comprises of smaller observations in terms of length, as generative models are computationally much more demanding when it comes to input size. The SentiGAN architecture is designed for generating text with specific sentiment types. It consists of k generators, each responsible for generating text with a specific sentiment label, and a discriminator. The generators use noise from a distribution Pz, such as a normal distribution, as input and are trained to minimize a penalty-based objective in order to produce text that can deceive the discriminator. On the other hand, the discriminator is trained to classify the text as either fake (generated by the generators) or real with one of the k sentiment types. In the experiments, SentiGAN is set up to generate text with two sentiment types (positive and negative), in our case we will have as many sentiment types as we have in teh LIAR dataset.

3.3 Hybrid Model for Detecting Misinformation

The variety of different models used for misinformation detection is wide. Different Deep Learning architectures were used separately. However, some researchers also decided to test hybrid models to increase the performance of individual models of one type. The example hybrid models consist of *Convolutional Recurrent Neural Networks*, *Convolutional Restricted Boltzman Machine*, *Ensemble-Based Fusion*, and *Long Short-Term Memory Density Mixture Model*.

3.3.1 Convolutional Recurrent Neural Network (CRNN)

Recently, applying CNN and RNN models in a hybrid way was proven to achieve better performance in some applications of Natural Language Processing, Speech Recognition or Time-Series Classification. It is believed that real-world data are structured sequences with spatial-temporal trends (Lin et al., 2019). These models combine a Convolutional Neural Network for visual features extraction with a Recurrent Neural Network for sequential connection between features. Such models have been effectively applied for fake news detection. For example, (Ruchansky et al., 2017) proposed the structure of a deep hybrid model with *Capture* and *Score* modules for fake news detection.

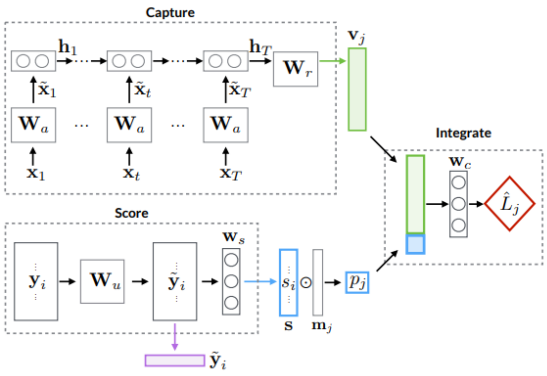

Figure 3: CSI model specification

In (Zhang et al., 2018b), researchers proposed deceptive review identification by RCNNs and demonstrated that the neural network approaches outperform the conventional techniques for all datasets.

3.3.2 Ensemble-Based Fusion (EBF)

In (Wang, 2017a), the first fake news detection benchmark dataset with speaker information was made. Wang used a hybrid model with speaker profiles data as the input. (Zhang et al., 2018a) analysed the problem of fake news detection on social media by identifying deceptive words used by online fake users harming society. Also, in

(Shu et al., 2018), fake news data repository FakeNewsNet was published for further analysis.

3.3.3 LSTM Density Mixture Model

Using only lexical features to detect fake news automatically was very popular methodology. Nonetheless, the hybrid deep neural network models started becoming more and more popular. Already mentioned (Ruchansky et al., 2017) analysed three different types of fake news:

- the text of an article, which our research will be focused on,
- the user response,
- the source on which users promote it.

They also analyzed how fake news affect public opinion. To do so, Ruchansky proposed a hybrid model combining all the characteristics to predict more accurately. Similarly, (Long et al., 2017) suggested a novel method with speaker profiles fed into attention-based LSTM architecture for fake news detection. LSTM-based models were proven to give better results for long sentences, whereas attention models provide weights representing importance of different words in context.

4 Datasets

We have decided to focus only on articles in which the average count of words exceeds 1000, since that would correspond to our task of detecting fake news as explained in section 2.1.

4.1 ISOT Fake News Dataset

The ISOT Fake News dataset contains both fake and real news, which were collected from Reuters.com source in case of real news, and different unknown sources for the fake news set in 2016 and 2017. In total there are 21417 articles labeled as real and 23481 labeled as fake. It is worth mentioning that these datasets are almost perfectly balanced. Both real and fake news have the following features: title, text, subject and date. The subject differs for each of datasets. For real dataset there are only two subjects: World-News and Politics-News. As for fake dataset there are: Government-News, Middle-east, US-news, left-news, politics and News. The figures 9 and 10 represent respectively word cloud and word count per article histogram obtained from the ISOT dataset.

4.2 Fake News Kaggle Dataset

The second dataset consists of news articles data. The dataset was published on Kaggle 5 years ago for the competition Fake News . It contains 18285 observations split into training and test set and is quite balanced with 43% fake news. The attributes of the dataset are presented as follows:

- `id`
- `title`
- `author`
- `text`
- `label`

The figures 11 and 12 represent respectively an example word cloud generated from the corpus of the dataset and its word count per article histogram.

4.3 LIAR dataset

In our case, the LIAR dataset (Wang, 2017b) is to be used solely for the purpose of testing trained models. It was manually labeled by PolitiFact and consists of short statements published in 2017. Each sentence was assigned one of the following labels: false, pants-fire, barely-true, half-true, mostly-true or true. There are in total 12836 cases. The dataset itself is a well-balanced set, having a range between 2000 to 2600 instances per class. Some of the attributes are:

- `label`
- `statement`
- `subject`
- `speaker`
- `speaker's job title`
- `party affiliation`

5 Methodology

In our case, we define a simple binary classification problem in which we will label articles as either real or fake. We will compare performance of a few different methods on this specific task.

5.1 Project Contributions

After further research of existing papers and taking into account your reviews, we have planned to work on a few unanswered questions in fake news detection tasks.

- How different preprocessing techniques influence the results of different architectures?
- Does keeping some stop words like yes or no influence the model's performance?
- Does a particular model trained on one political fake news dataset generalise to other datasets of a similar topic? Do models usually overfit, or are able to recognise patterns in fake news data?

5.2 Preprocessing

In order to prepare the data for our model we need to apply some preprocessing techniques such as:

- Using regular expressions to remove: punctuations, special symbols ($, #, &), digits, URLs, wide spaces, single characters
- Removing stop words
- Lower casing
- Lemmatization using two different libraries (SpaCy, NLTK)
- Tokenization suited for a particular architecture
- Article padding

We might have to also consider additional preprocessing steps depending on which type of model we will decide to use.

5.3 Proposed Architectures

First method will involve combination of word embedding using word2vec and LSTM for classification task. Further customizations include different loss functions like cross-entropy loss or log-likelihood and other hyperparameters.

The next architectures that we want to apply for the given problem are hybrid models, which were proven to result in accurate classification. We will start with simple CNN+RNN models that extract both shape-related and sequential features of the articles' texts. At this point, ensemble models with lighter architectures may also be introduced to provide variety of approaches for the given task.

Lastly, we are planning to test state-of-the-art attention-based architectures with pre-trained text representation model like BERT to extract text features, (Matsumoto et al., 2021).

6 Results

6.1 Word2vec with LSTM

We have tested a mix of word2vec model for word embeddings with LSTM neural network. We applied all preprocessing techniques mentioned in section 5.2. The optimizer used for training is Adam. The model consists of:

1. embedding layer obtained with word2vec
2. 32 LSTM units
3. dropout layers

The model's accuracy quickly converged to around 99% in just 5 training epochs for both ISOT and FakeNews datasets. Trained models were then tested on the LIAR dataset. We relabeled data so that classes 'false', 'barely-true' and 'pants-fire' is considered as fake news and the others as true news. We present the results in Table 1.

Dataset	Val. accuracy	F1 Score	LIAR accuracy	LIAR F1 Score
ISOT	0.9991	0.9992	0.4723	0.2189
FakeNews	0.9495	0.9565	0.5391	0.6894

Table 1: Results for word2vec with LSTM model

High accuracy for training sets did not lead to any substantial results in case of tests performed on external dataset. While both of them were based on politic topics, it is most likely that training datasets were not generic and couldn't be solely used for misinformation classification.

6.2 BERT Results

We have tested one of the most ground-breaking model in Natural Language Processing, Bidirectional Encoder Representations from Transformers (BERT). Because of the complexity of the model, we limited our experiments to a few number of epochs in training.

We used pretrained `BertForSequenceClassification` with `BertTokenizer` implemented in `Transformers` library in Python. The

chosen architecture consisted of 124 layers, 1024 hidden layers, and 16 heads. The weight of the model itself was about 450 megabytes. The model was pretrained on uncased vocabulary. The model was trained on a machine with Intel Core i7-11700KF CPU and GeForce RTX 3060 GPU with 12 gigabytes of VRAM.

For the training, we used the following hyperparameters and charcteristics:

- Maximum length of tokens - 512 (a popular length for pretrained transformers)
- Fixed seed for both data loaders and the model
- 80-20 training-validation split
- AdamW optimizer with 10^{-5} learning rate
- 5 epochs of training

6.2.1 Different Preprocessing Techniques

The research focused on the analysis how different preprocessing techniques influene the performance of the BERT model. Three different scenarios were tested:

- only regular expressions
- regular expressions and removing stop words
- regular expressions, removing stop words lemmatisation

Figure 4, 5, 6 represent the learning curves for all scenarios for validation loss, accuracy, and F1-score.

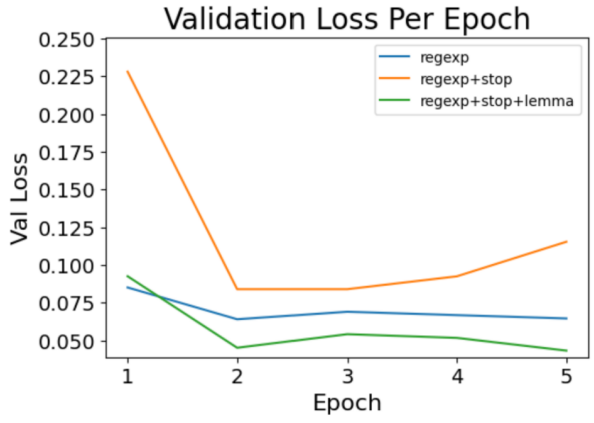

Figure 4: Validation loss of a model for three different scenarios

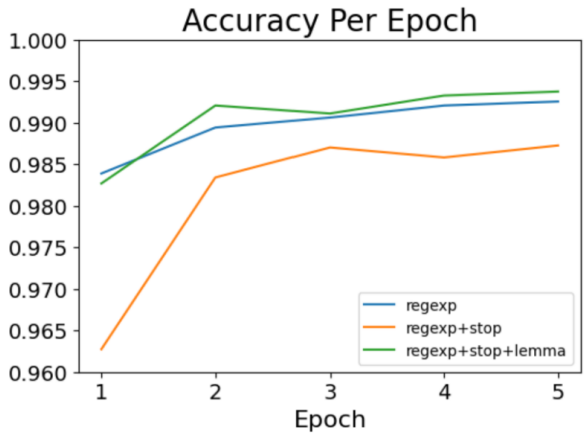

Figure 5: Accuracy of a model for three different scenarios

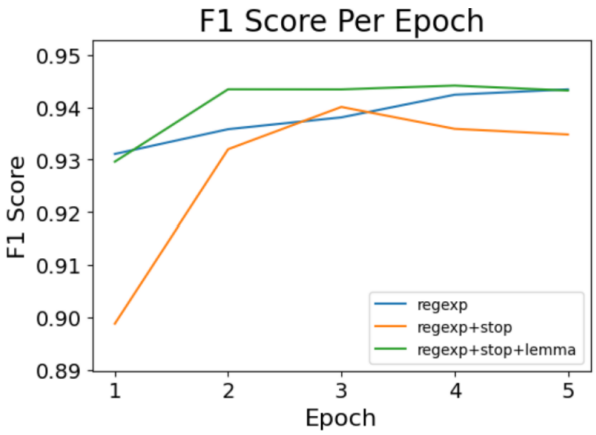

Figure 6: F1-score of a model for three different scenarios

It occurred that the BERT model with the largest number of preprocessing techniques obtained the best results. However, the increase in the performance was not significant. What is more, choosing different stop words did not influence the performance of the model.

6.2.2 Testing a Model On Similar Datasets

We trained the BERT model on a Kaggle dataset and then tested its performance on the ISOT Dataset without any fine-tuning. Similarly, we applied different sets of preprocessing techniques for testing purposes. Again, it occurred that the scenario with the largest number of techniques obtained the best results. But this time, the difference in the performance is significant.

6.2.3 Training BERT on LIAR Dataset

As a point of reference, we also trained our BERT model on LIAR dataset (Wang, 2017b) with using

Preprocessing Techinques	Accuracy	F1 Score	Validation Loss
Regular Expressions	0.28502	0.34156	8.56201
Reg. Exp. + Stopwords	0.34392	0.45906	7.18088
Reg. Exp. + Lemmatisation	0.32753	0.41701	8.20974
Reg. Exp. + Stopwords + Lemma.	**0.41517**	**0.57178**	**5.92430**

Figure 7: Performance on a similar dataset without fine-tuning

all preprocessing techniques. After 5 epochs, we obtained the following results:

- Accuracy: 0.61510
- F1: 0.62789
- Validation Loss: 0.65884

One year after the publication of the LIAR dataset, other researchers published an enhanced version of the dataset. The LIAR-PLUS dataset (Alhindi et al., 2018) consisted of justifications to the existing annotations. This gives models much more information to increase the performance. Figure 8 shows the results obtained for the LIAR-PLUS dataset by the researchers.

Testing Set	Binary Classification	6 Class Classification
LIAR-PLUS Test set	77.2%	37.4%

Figure 8: Results obtained by the researchers on the LIAR-PLUS dataset (Alhindi et al., 2018)

We are planning to test the enhanced input during the second project. Additionally, the hyperparameters of the architecture will be tuned to increase the performance.

7 Further Research

We plan to continue working on fake news classification task with the following extensions.

- multi-class classification with other types of misinformation or the degree of certainty that a piece of news is fake (for example, 0-5 scale)
- datasets with extended input, for example authors' data, context, source, category
- providing analysis how the choice of inputs influence the models' performance
- different methods used in state-of-the-art approaches that can be improved
- application of generative models: SentiGAN and CatGAN

Appendix

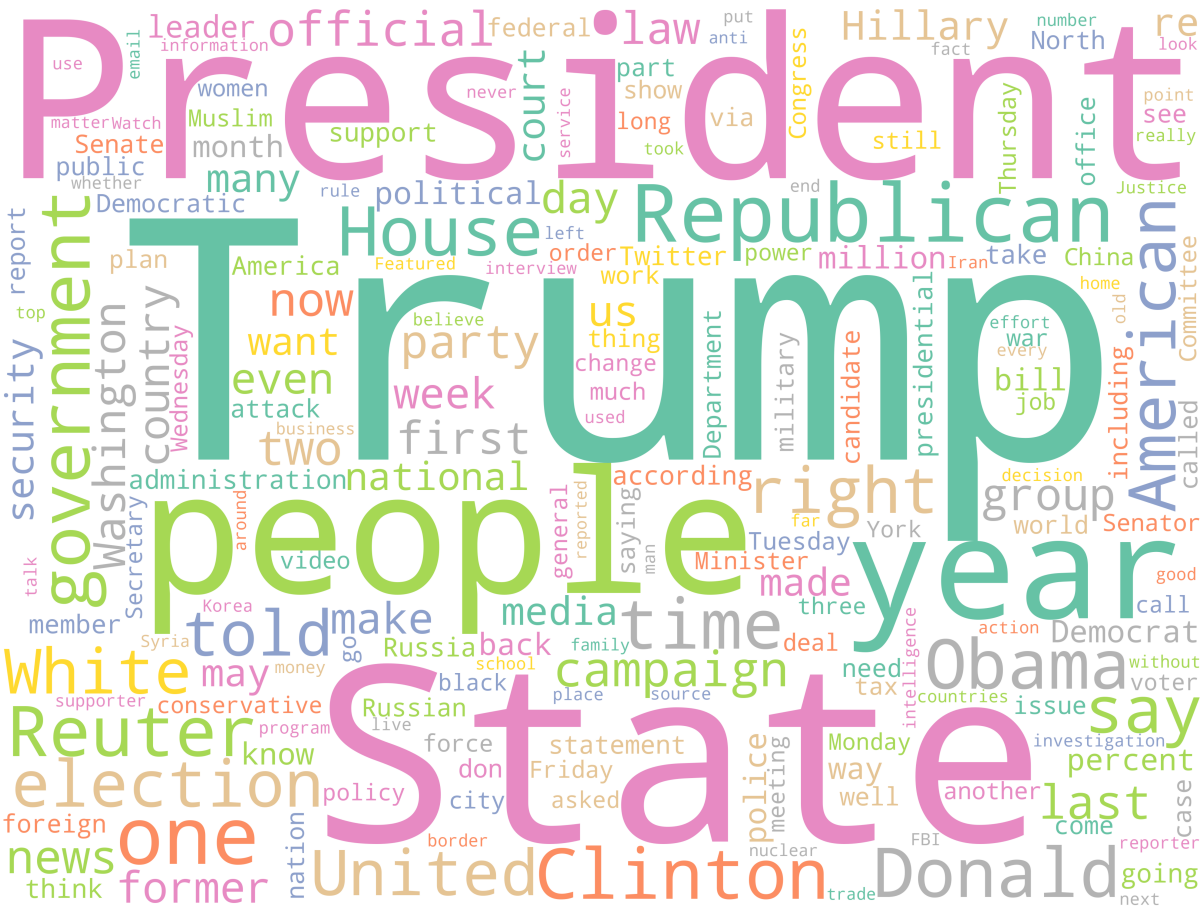

Figure 9: Word cloud for ISOT Fake News Dataset

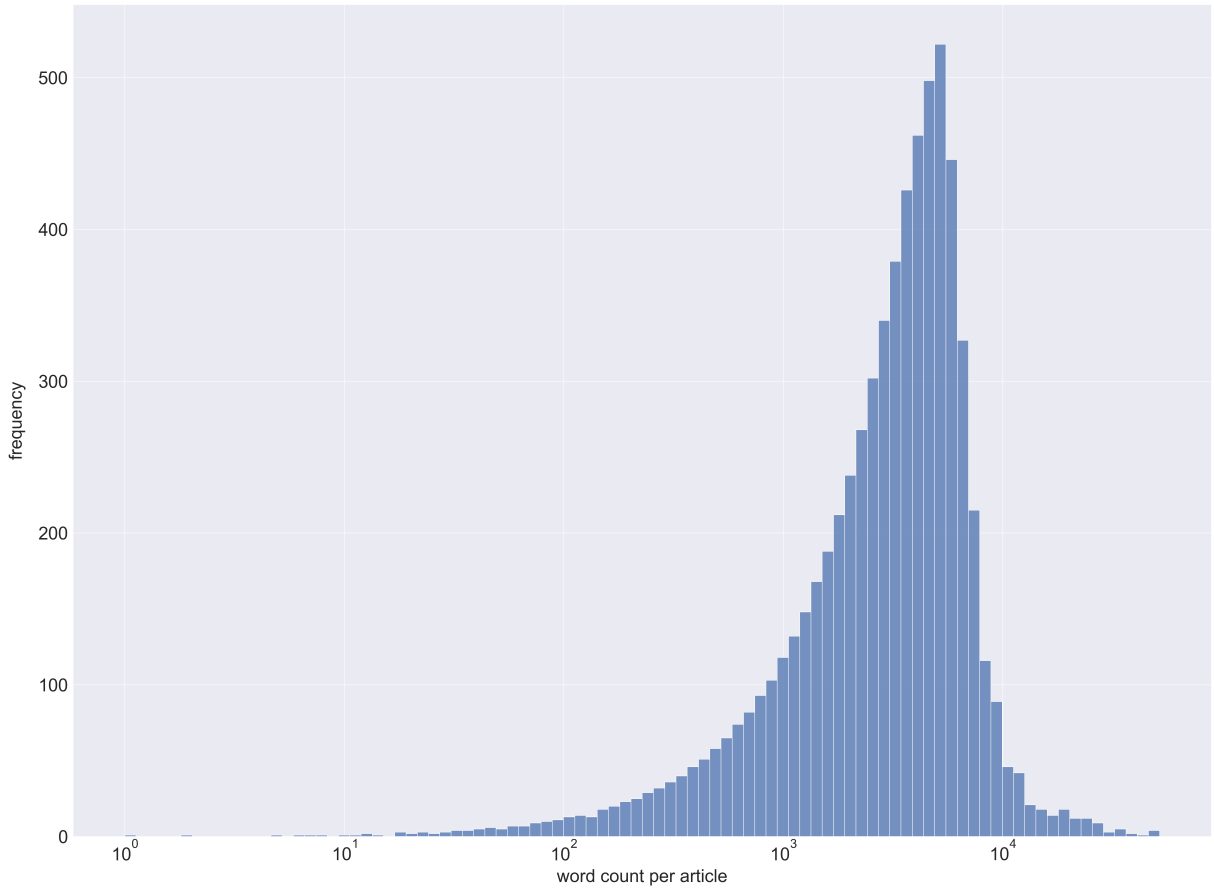

Figure 10: Word count of articles histogram for ISOT Fake News Dataset

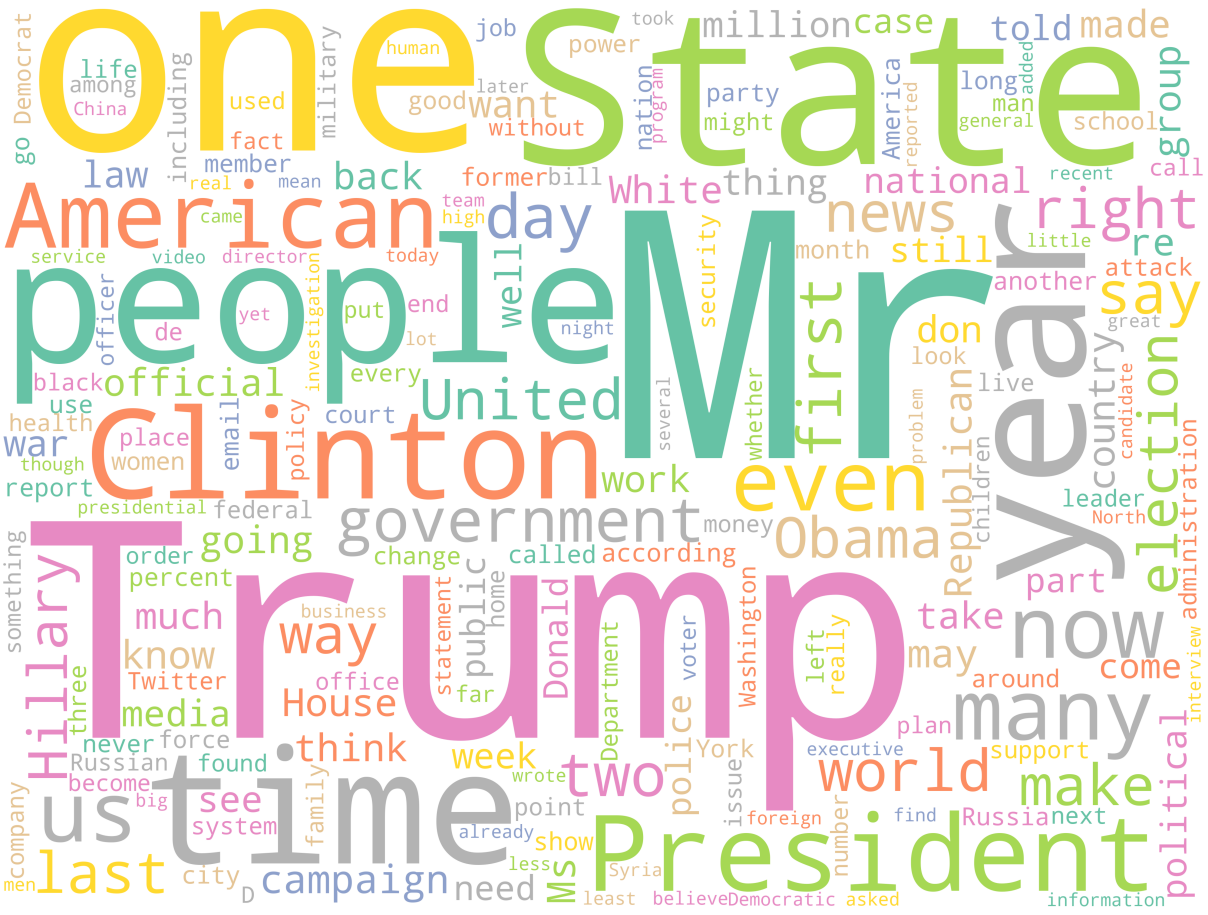

Figure 11: Word cloud for Kaggle Fake News Dataset

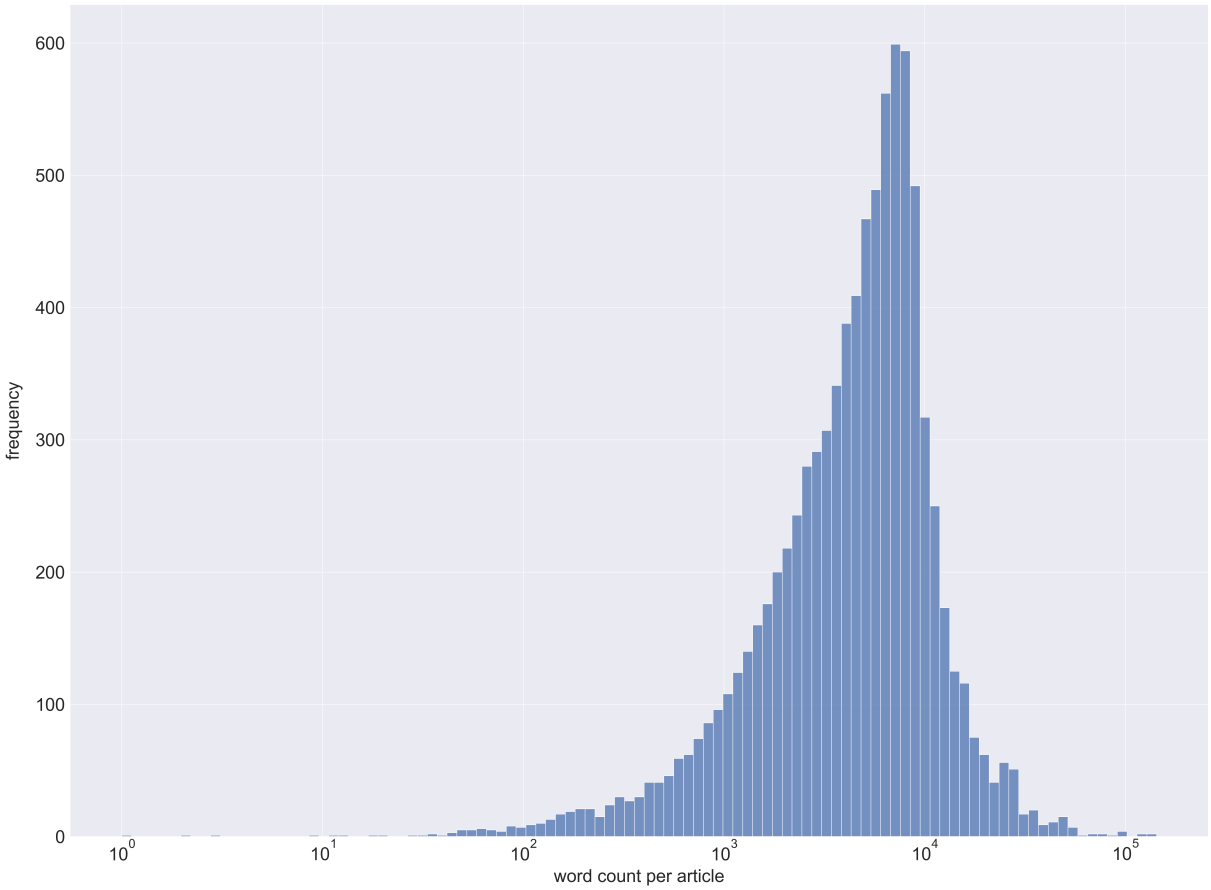

Figure 12: Word count of articles histogram for Kaggle Fake News Dataset

References

[Alhindi et al.2018] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium, November. Association for Computational Linguistics.

[Alkhodair et al.2020] Sarah A. Alkhodair, Steven H.H. Ding, Benjamin C.M. Fung, and Junqiang Liu. 2020. Detecting breaking news rumors of emerging topics in social media. *Information Processing Management*, 57(2):102018.

[Chen et al.2017a] Tong Chen, Lin Wu, Xue Li, Jun Zhang, Hongzhi Yin, and Yang Wang. 2017a. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection.

[Chen et al.2017b] Yi-Chin Chen, Zhao-Yang Liu, and Hung-Yu Kao. 2017b. IKM at SemEval-2017 task 8: Convolutional neural networks for stance detection and rumor verification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 465–469, Vancouver, Canada, August. Association for Computational Linguistics.

[Islam et al.2020] Md. Rafiqul Islam, S. Liu, Xianzhi Wang, and Guandong Xu. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10.

[Jacovi et al.2018] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018. Understanding convolutional neural networks for text classification.

[Ke Wang2018] Xiaojun Wan Ke Wang. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. July.

[Lin et al.2019] Xiang Lin, Xiangwen Liao, Tong Xu, Wenjing Pian, and Kam-Fai Wong. 2019. Rumor detection with hierarchical recurrent convolutional neural network. In *NLPCC*.

[Long et al.2017] Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *IJCNLP*.

[Ma et al.2018] 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, July. Association for Computational Linguistics.

[Ma et al.2019] Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The World Wide Web Conference*, WWW '19, page 3049–3055, New York, NY, USA. Association for Computing Machinery.

[Matsumoto et al.2021] Hayato Matsumoto, Soh Yoshida, and Mitsuji Muneyasu. 2021. Propagation-based fake news detection using graph neural networks with transformer. pages 19–20, 10.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

[Qian et al.2018] Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. 2018. Neural user response generator: Fake news detection with collective user intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 3834–3840. AAAI Press.

[Ruchansky et al.2017] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A hybrid deep model for fake news. *CoRR*, abs/1703.06959.

[Shu et al.2018] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *CoRR*, abs/1809.01286.

[Wang2017a] William Yang Wang. 2017a. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *CoRR*, abs/1705.00648.

[Wang2017b] William Yang Wang. 2017b. "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, July. Association for Computational Linguistics.

[Wei et al.2018] Lei Wei, Donghuai Gao, and Cheng Luo. 2018. False data injection attacks detection with deep belief networks in smart grid. In *2018 Chinese Automation Congress (CAC)*, pages 2621–2625.

[Wu et al.2017] Liang Wu, Jundong Li, Xia Hu, and Huan Liu, 2017. *Gleaning Wisdom from the Past: Early Detection of Emerging Rumors in Social Media*, pages 99–107. 06.

[Yang et al.2018] Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S. Yu. 2018. Ti-cnn: Convolutional neural networks for fake news detection.

[Zhang et al.2018a] Jiawei Zhang, Limeng Cui, Yanjie Fu, and Fisher B. Gouza. 2018a. Fake news detection with deep diffusive network model. *CoRR*, abs/1805.08751.

[Zhang et al.2018b] Wen Zhang, Yuhang Du, Taketoshi Yoshida, and Qing Wang. 2018b. Dri-rcnn: An approach to deceptive review identification using recurrent convolutional neural network. *Information Processing Management*, 54(4):576–592.

[Zubiaga et al.2016] Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2438–2448, Osaka, Japan, December. The COLING 2016 Organizing Committee.