Fake News Identification Based on Sentiment and Frequency Analysis

Jozef Kapusta[1,2(✉)], Ľubomír Benko[1], and Michal Munk[1]

[1] Department of Informatics, Constantine the Philosopher University in Nitra, Nitra, Slovak Republic
{jkapusta, lbenko, mmunk}@ukf.sk
[2] Institute of Computer Science, Pedagogical University of Cracow, Cracow, Poland
jkapusta@up.krakow.pl

Abstract. The advent of social networks has changed how can be the thinking of the population influenced. Although the spreading of false information or false messages for personal or political benefit is certainly nothing new, current trends such as social media enable every individual to create false information easier than ever with the spread compared to the leading news portals. Fake news detection has recently attracted growing interest from the general public and researchers. The paper aims to compare basic text characteristics of fake and real news article types. We analysed two datasets that contained a total of 28 870 articles. The results were validated using the third data set consisting of 402 articles. The most important finding is the statistically significant difference in the news sentiment where it has been shown that fake news articles have a more negative sentiment. Also, an interesting result was the difference of average words per sentence. Finding statistically significant differences in individual text characteristics is a piece of important information for the future fake news classifier in terms of selecting the appropriate attributes for classification.

Keywords: Fake news identification · Text mining · Sentiment analysis · Frequency analysis

1 Introduction

Fake News is currently the biggest bugbear of the developed world. Alongside increasing use of social networks, especially for communication, we observe the high increase in the distribution of false news, hoaxes and other half-truths in periodicals, as well as in society. People perceive social networks in particular as a space for expressing their opinions, but also as a space that brings together people with similar opinions.

Although the spreading of false information or false messages for personal or political benefit is certainly nothing new, current trends such as social media enable every individual to create false information easier than ever with the spread compared to the leading news portals [1]. Importantly, while the spread of false information is simple, the correction of the record can be much more difficult.

Most of use just scans social networks posts, so there is no time to confront the source of information, compared to the face-to-face dialogue. People simply register

M. Serrhini et al. (Eds.): EMENA-ISTL 2019, LAIS 7, pp. 400–409, 2020.
https://doi.org/10.1007/978-3-030-36778-7_44

the message. The general awareness that the media have the possibility of mutual cross-checking merely enhances this problem. It is convenient to slip into the belief that the number of people who have the opportunity to verify the validity of information has certainly done so.

The paper analyses the available datasets of fake and real news. Basic analysis of these datasets was focused on the text content of the articles. We analysed the average number of words in sentences, number of stop words as well as basic classification of the sentiment of the examined news. The aim of our paper is to find out whether there are statistically significant differences in basic text characteristics between the fake news articles comparing with the real news articles.

In the second chapter of the article, we summarize the current trends in fake news detection and fight against them. The third chapter is focused on the description of the examined dataset and the description of the data preparation method, which we applied to it. In the fourth chapter, we present the results of our analyses, and the discussion of the results is the content of the last chapter.

2 Related Work

An experiment within the research of fake news studied, how people adapt their opinion if the information which formed the opinion, were incorrect. Research has shown that "correction" of opinion depends on the cognitive abilities of an individual. The results indicate that the correction of incorrect information does not always lead to a change of opinion [2]. In another research, 307,738 tweets with 30 fake and 30 real messages were analysed. The findings revealed that fake news tweets were mostly generated by regular users and often included a link to untrusted news sites. The results also showed that tweets about true news spread widely and rapidly, while tweets with fake news have been modified several times in the process of dissemination and therefore they spread slower [3]. There are several other similar studies on the negative impact of fake news as well as on the formation of opinions depending on the fake news [3, 4].

The seriousness of the situation illustrates the fact that the fight against the spread of fake news and development of effective strategy in this field is among the highest priorities of the EU. The European Commission has conducted a questionnaire survey on fake news. According to [5], more than 37% of respondents from EU countries are confronted with false reports on a daily basis. The key findings from the research are damaging society in areas such as political affairs or ethnic minorities. Research highlights the extent to which fake news is present in EU countries.

As the topic of fake news is very current, many researchers try to overcome the issues of identifying fake news articles in the number of real news articles. Xu et al. [6] have characterized hundreds of popular fake and real news measured by shares, reactions, and comments on Facebook from two perspectives: Web sites and content. The presented analysis concludes that there are differences between fake and real news publisher's web sites in the behavior of user registration. Also, the fake news tends to disappear from the web after a certain amount of time. The authors applied exploration of document similarity with the term and word vectors for predicting fake and real news. Braşoveanu and Andonie [7] introduced a novel approach to fake news detection combining machine learning, semantics and natural language processing. The authors

used relational features like sentiment, entities or facts extracted directly from the text. The authors concluded that using the relational features together with syntactic features, it is possible to beat the baselines even without using advanced architectures. The experiment showed that consideration of relational features can lead to an increase in the accuracy of the most classifiers. Saikh et al. [8] correlated the Fake News Challenge Stage 1 (FNC-1) dataset that introduced the benchmark FNC stage-1: stance detection task with Textual Entailment. The stance detection task could be an effective first step towards building a robust fact-checking system. The proposed model outperformed the state-of-the-art system in FNC and F1 score, and F1 score of Agree class by the third Deep Learning model i.e. the model augmented with Textual Entailment features. The authors in [9] have focused on creating a model for fake news detection using the Python programming language. The authors used the Naive Bayes algorithm with two forms of tokenization- CountVectorizer, and TfidfVectorizer. The results showed that the CountVectorizer was more successful classification method since it achieved the accuracy of 89.3% of the news correctly classified.

3 Materials and Methods

For our analysis we created a dataset that consists of merging many existing datasets:

1. Dataset of real news was created from articles analysed during three months[1] that were validated using https://mediabiasfactcheck.com. The dataset contained 15 707 articles.
2. Dataset of fake news was created[2] based on the text analysis of 244 web pages marked as "bullshit" from BS Detector Chrome Extension [10] by Daniel Sieradski. The important fact is that these articles were published in the same period (October – December 2016) as the articles in real news dataset. The fake news dataset contained 12 761 articles.
3. Dataset KaiDMML, the dataset of fake and real news was taken over[3] and processed based on [11]. The dataset was relatively less extensive, including articles also from the fake news group (205 articles) as well as real news (197 articles). Despite the small number of articles in the dataset, we have taken it as the best-created one and in our analysis, we used it to verify the facts found in the first two datasets.

The first two datasets were available on the server kaggle.com. We chose them because both datasets were already used in the analysis focused on fake news classifications and similar it was with the third dataset [12]. It were therefore datasets that were verified by other researchers. The following basic methods were applied to the created dataset:

1. For each article, the average number of words in the sentence was calculated.
2. So-called stop words were isolated from articles and the average number of words in the sentence without stop words was calculated. The stop words identification

[1] https://www.kaggle.com/anthonyc1/gathering-real-news-for-oct-dec-2016.

[2] https://www.kaggle.com/mrisdal/fake-news.

[3] https://github.com/KaiDMML/FakeNewsNet.

was done using word comparison from the article with a list of stop words for the English language.
3. The sentiment rate for each article was calculated. This sentiment rate was done using the most basic method of difference between the number of positive and negative words in the article. It was a relative abundance so the difference was still divided by the number of words in the analysed article.

4 Results of Sentiment and Frequency Analysis

The first analysis is focused on the basic overview of the average number of sentences in articles, number of words and the average number of words in the sentence. These were calculated using a separate algorithm where we used the NLTK library dedicated for analysis of natural language processing using programming language Python. Using the library's functions we made the basic sentence tokenization and within sentences verbal tokenization. Quantities such as the average number of sentences, number of words in the article and the average number of words in the sentence were calculated for each article. Similarly, we calculated the listed variables after removing the stop words from the examined articles.

The analysis of variance was used for testing the differences between independent samples (fake: 0/1) in count sentences, count words per sentence, count words without stop words per sentence and in the measure of sentiment. The null hypotheses state that there is no statistically significant difference in the number of sentences/words and in the measure of sentiment between fake and real news, i.e. classifying the news as fake do not depend on the number of sentences/words and the measure of sentiments.

Based on ANOVA results (Table 1), we reject the null hypotheses at the 1% significant level, i.e. it was proven a statistically significant difference in the count sentences between fake and real news over all datasets (Table 1a) as well as in the KaiDMML validation dataset (Table 1b) (Fig. 1).

Table 1. Univariate results for count sentences (a) All Datasets (b) KaiDMML

	df	SS	MS	F	p
All Datasets: count_sentences					
Intercept	1	38950472.0	38950472.0	15270.32	0.0000
Fake	1	2134151.0	2134151.0	836.68	0.0000
Error	28869	73637057.0	2551.0		
Total	28870	75771208.0			
KaiDMML: count_sentences					
Intercept	1	218778.7	218778.7	335.82	0.0000
Fake	1	5452.7	5452.7	8.37	0.0040
Error	401	261243.2	651.5		
Total	402	266695.9			

df - degrees of freedom, SS - sum of squares, MS - mean square, F- test statistic, p - probability value

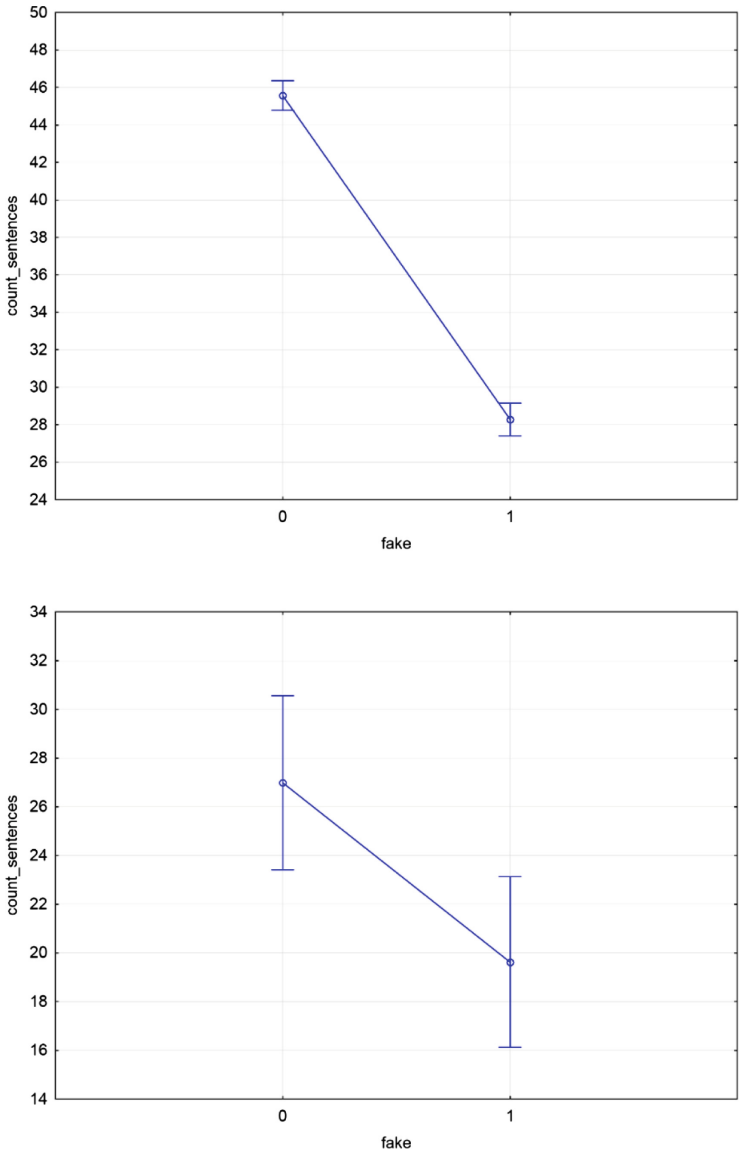

Fig. 1. Mean plot: point and interval estimates for count sentences (a) All Datasets (b) KaiDMML

The results show that there were fewer sentences in the fake news articles (approximately 28 sentences per one article) in comparison to the real news articles. If we take into account only the last third dataset, i.e. dataset that belongs to verified and recommended datasets, then the result is similar, although the amount of sentences per article is smaller (approximately 21 sentences per one fake news article).

Another view on the examined data was the comparison of the average number of words in the sentence. We also examined the average number of words after removing so-called stop words. Similarly, in the case of count words, based on ANOVA results (Table 2), we reject the null hypotheses at the 0.1% significant level, i.e. a statistically significant difference in the count words per sentence (Table 2a) was proven as well as in the count words without stop words per sentence (Table 2b) between fake and real news. Similar results were also achieved in the KaiDMML validation dataset (Fig. 2).

Table 2. Univariate results for count words (a) count words per sentence (b) count words without stop words per sentence

	df	SS	MS	F	p
All Datasets: words_per_sentence					
Intercept	1	20911584.0	20911584.0	185292.30	0.0000
Fake	1	29887.0	29887.0	264.80	0.0000
Error	28869	3258077.0	113.0		
Total	28870	3287965.0			
All Groups: words_wo_stopw_per_sentence					
Intercept	1	9222142.0	9222142.0	145087.10	0.0000
Fake	1	28400.0	28400.0	446.80	0.0000
Error	401	1834995.0	64.0		
Total	402	1863395.0			

df - degrees of freedom, SS - sum of squares, MS - mean square, F- test statistic, p - probability value

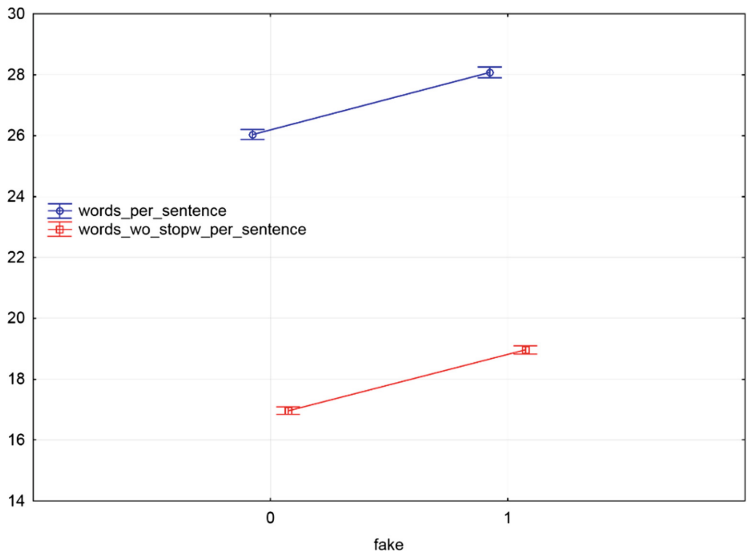

Fig. 2. Mean plot: point and interval estimates for count words

It is obvious that the average number of words in sentences after the stop words removal was clearly smaller. It is interesting that by the experiment we found out that there is a statistically significant difference in the average number of words in sentence between the fake news and real news articles. This significant difference is observed whether after the removal of stop words or without this modification. Based on the results we can state that the fake news articles use more complex sentences, i.e. sentences that contain more words.

A separate view was the research of positivity or negativity of the examined articles, the so-called sentiment of articles. The sentiment analysis was done only using the basic method of determining the sentiment since the sentiment analysis itself was not one of the main objectives of our research. However to verify general awareness of fake news, i.e. that fake news articles are negative oriented, was done this analysis. Similarly, in the case of measure of sentiment, based on ANOVA results (Table 3), we reject the null hypotheses at the 0.1% significant level, i.e. a statistically significant difference in sentiment between fake and real news was proven over all datasets (Table 3a) as well as in the KaiDMML validation dataset (Table 3b) (Fig. 3).

Table 3. Univariate results for sentiment (a) All Datasets (b) KaiDMML

	df	SS	MS	F	p
All Datasets: sentiment					
Intercept	1	0.478	0.478	1277.72	0.0000
Fake	1	0.013	0.013	34.13	0.0000
Error	28869	10.811	0.000		
Total	28870	10.824			
KaiDMML: sentiment					
Intercept	1	0.022	0.022	62.43	0.0000
Fake	1	0.004	0.004	11.68	0.0007
Error	401	0.139	0.000		
Total	402	0.143			

df - degrees of freedom, SS - sum of squares, MS - mean square, F- test statistic, p - probability value

Based on the results it can be seen that the sentiment of all articles is rather negative. Also, the sentiment of each article is a relative small value. This is due to the simple calculation of sentiment that is calculated as the division of the difference between positive and negative words by the number of all words in the article.

Based on the results was identified a statistically significant difference in the sentiment of fake news articles compared to the real news articles. This difference was demonstrated as for the whole dataset, as well as for the control dataset KaiDMML. Fake news articles are statistically significantly more negative.

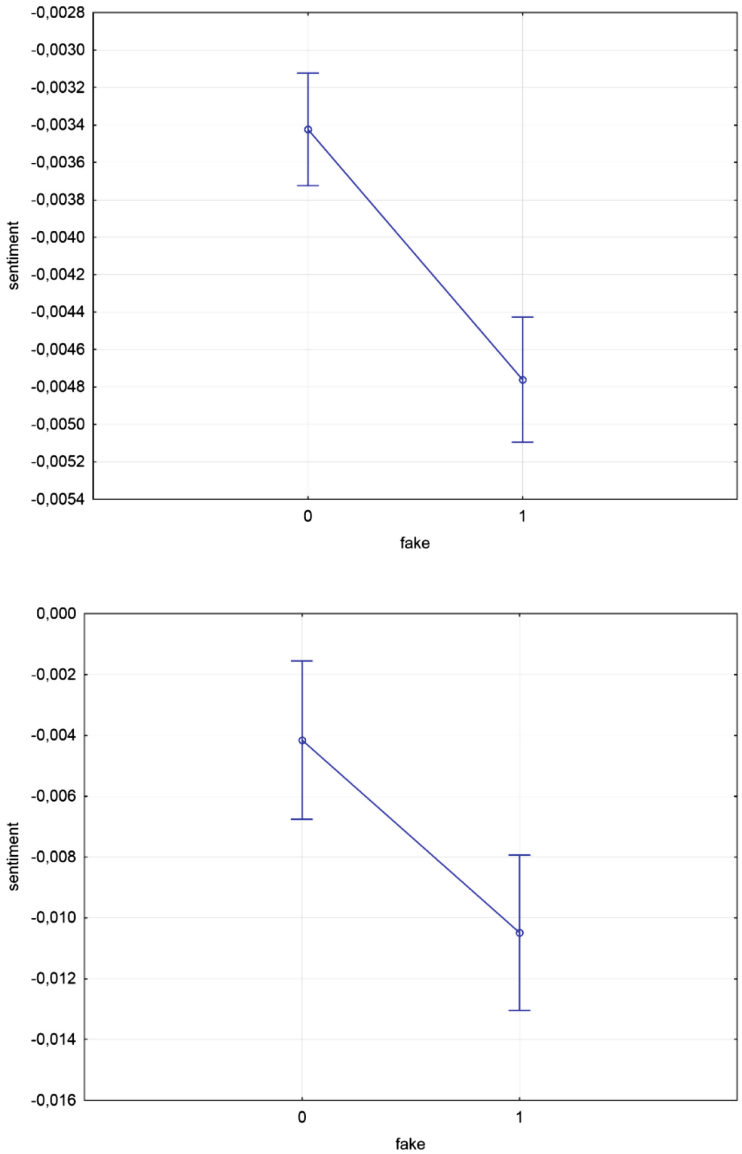

Fig. 3. Mean plot: point and interval estimates for sentiment (a) All Datasets (b) KaiDMML

5 Discussion and Conclusion

In this paper, we analysed the fake and real news dataset that was created from three existing datasets. All three used datasets are currently freely available datasets. Their disadvantage is that they contain articles from the year 2016. It is possible that the style of fake news articles is changing and our results may no longer reflect the latest trends

in writing fake news articles. However, they are verified and frequently analysed datasets and for this reason, they have been selected.

The second problem of the datasets is their creation. Especially the first used dataset was created exclusively by the fake news classifier. It, therefore, contains articles that certainly have not been evaluated by a human. As a result of the fake news classifier, it contains articles that were clearly determined as fake news and probably the dataset does not contain borderline articles.

Our analysis was focused on the basic characteristics of text in articles of the examined datasets. Among the most important findings was the confirmation of statistically significant difference in the article sentiment and it turned out that fake news articles had a more negative sentiment. The result is interesting mainly because of the relatively simple method of classification of the sentiment. If we were able to verify a statistically significant difference using the basic sentiment classification method, we can assume that using a more sophisticated classification method will confirm this result and even the differences in sentiment will be more pronounced.

An interesting result was also the finding of a statistically significant difference in the average number of words per sentence. This was identified also for articles where stop words were removed. The result was a surprise for us, we assumed that the fake news articles would be simpler, more accurate, i.e. with a lower average number of words per sentence. The results have shown that fake news are probably trying to mislead the reader with their more complicated, more descriptive style.

Finding statistically significant differences in individual textual characteristics is an important piece of information for the future fake news classifier in terms of selecting appropriate attributes for classification.

Acknowledgment. This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and of Slovak Academy of Sciences under the contract VEGA-1/0776/18.

This publication was supported by the Operational Program: Research and Innovation project "Fake news on the Internet - identification, content analysis, emotions", co-funded by the European Regional Development Fund.

References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. J. Econ. Perspect. **31**, 211–236 (2017). https://doi.org/10.1257/jep.31.2.211
2. De Keersmaecker, J., Roets, A.: 'Fake news': incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. Intelligence **65**, 107–110 (2017). https://doi.org/10.1016/J.INTELL.2017.10.005
3. Jang, S.M., Geng, T., Queenie Li, J.-Y., Xia, R., Huang, C.-T., Kim, H., Tang, J.: A computational approach for examining the roots and spreading patterns of fake news: evolution tree analysis. Comput. Hum. Behav. **84**, 103–113 (2018). https://doi.org/10.1016/J.CHB.2018.02.032
4. Brigida, M., Pratt, W.R.: Fake news. North Am. J. Econ. Finance **42**, 564–573 (2017). https://doi.org/10.1016/J.NAJEF.2017.08.012

5. Eurobarometer 464 – Fake news and disinformation online. http://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/ResultDoc/download/DocumentKy/82798
6. Xu, K., Wang, F., Wang, H., Yang, B.: A first step towards combating fake news over online social media. Presented at the June (2018). https://doi.org/10.1007/978-3-319-94268-1_43
7. Brașoveanu, A.M.P., Andonie, R.: Semantic fake news detection: a machine learning perspective. Presented at the June (2019). https://doi.org/10.1007/978-3-030-20521-8_54
8. Saikh, T., Anand, A., Ekbal, A., Bhattacharyya, P.: A novel approach towards fake news detection: deep learning augmented with textual entailment features. Presented at the June (2019). https://doi.org/10.1007/978-3-030-23281-8_30
9. Agudelo, G.E.R., Parra, O.J.S., Velandia, J.B.: Raising a model for fake news detection using machine learning in Python. Presented at the October (2018). https://doi.org/10.1007/978-3-030-02131-3_52
10. Jane Wakefield: Fake news detector plug-in developed - BBC News. https://www.bbc.com/news/technology-38181158
11. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: FakeNewsNet: a data repository with news content, social context and spatialtemporal information for studying fake news on social media (2018)
12. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media. ACM SIGKDD Explor. Newsl. **19**, 22–36 (2017). https://doi.org/10.1145/3137597.3137600