Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2019) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

Comparison of fake and real news based on morphological analysis

Jozef Kapusta[a,b,]*, Petr Hájek[c], Michal Munk[a], Ľubomír Benko[a]

[a]*Department of Informatics, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia*
[b]*Institute of Computer Science, Pedagogical University of Cracow, ul. Podchorążych 2, 30-084 Kraków, Poland*
[c]*Institute of System Engineering and Informatics, University of Pardubice, Studentska 95, 532 10 Pardubice, Czech Republic*

Abstract

Easy access to information results in the phenomenon of false news spreading intentionally through social networks to manipulate people's opinions. Fake news detection has recently attracted growing interest from the general public and researchers. The paper deals with the morphological analysis of two datasets containing 28 870 news articles. The results were verified using the third dataset consisting of 402 news articles. The analysis of the datasets was carried out using lemmatization and POS tagging. The morphological analysis as a process of classifying the words into grammatical-semantic classes and assigning grammatical categories to these words. Individual words from articles were annotated and statistically significant differences were examined between the classes found in fake and real news articles. The results of the analysis show that statistically significant differences are mainly in the verbs and nouns word classes. Finding statistically significant differences in individual categories of word classes is an important piece of information for the future fake news classifier in terms of selecting appropriate variables for the classification.

Keywords: fake news identification, text mining, natural language processing, post-editing, POS tagging, morphological analysis

* Corresponding author. Tel.: +421 37 6408 675; fax: +421 37 6408 556.
E-mail address: jkapusta@ukf.sk

1. Introduction

The biggest problem of nowadays online media newspapers is mostly fake news articles. There is an increase in the distribution of false news, hoaxes and other half-truths in society. The spread of fake news is done not only in the virtual world (social media, online media, etc.) but also from person to person. Fake news is a big thread because this way they can affect many people around the world every day. Niklewicz [1] views social media as a dominant source of information for significant parts of the societies. Niklewicz focused on those aspects of social media that negatively affect the public debate, such as is the spreading of fake news and proposed that social media platforms should be considered media companies and that they should be regulated by modified versions of existing press laws, adapted to suit the new technology.

This paper is focused on the available datasets of fake and real news. The experiment is oriented on the linguistic side of the text using the morphological analysis of the news content. The morphological analysis is the basic tool to examine the natural language. It contains word-specific and form characteristics of words in context. The result is a set of tags that describe the grammatical categories of a given word form, especially the morpheme and morphological case.

This paper aims to find out whether there are statistically significant differences in the usage of word-specific and word-form characteristics between real news and articles classified as fake news. This analysis can help to identify whether there is a difference in the usage, or in preferring some word types and word shape characteristics when creating fake news.

The paper is divided into six sections. The second section is focused on the summary of the related work in the field of fake news identification. The third section describes the used datasets and methods of the dataset preparation for the analysis. The fourth section is focused on the results of the analysis. Subsequently, the last two sections contain the discussion and conclusion of the experiment.

2. Related Work

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. [2] 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 the exploration of document similarity with the term and word vectors for predicting fake and real news. Braşoveanu and Andonie [3] 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. [4] 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 [5] 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 a more successful classification method since it achieved the accuracy of 89.3% of the news correctly classified.

Ahmed et al. [6] proposed a fake news detection model that uses n-gram analysis and machine learning techniques. The authors have experimented with two different features of extraction techniques and six different machine learning techniques. The highest accuracy was obtained for a model using unigram features and Linear SVM classifier (it achieved an accuracy score of 92%). Zhuk et al. [7] have used machine learning to identify the main features of fake news. The proposed model has learned to analyze how the text is written and to determine whether it has the evaluative vocabulary, author's judgments, and words with strong connotations or obscene expressions. Shu et al. [8] focused on fake news detection using a proposed tool FakeNewsTracker. It is a system for fake news detection and understanding. The detection was done using deep learning-based solutions with linguistic and social engagement features. The proposed tool offers also a visualization option to better interpret the results. Bhutani et al. [9] focused on the detection of fake news that involves sentiment as a feature to improve the accuracy of detection. The performance of the novel solution was tested using three different datasets. The results showed improvement against the other text processing techniques. Dey et al. [10] analyzed the US Presidential Election for news deception and identifying the hidden bias of the author. The authors have done a linguistic analysis of tweets and applied the k-nearest neighbor algorithm. This way the authors divided fake news from real news.

3. Materials and Methods

For the need of the experiment, it was necessary to create a dataset. The dataset consists of many existing datasets merged into the examined dataset:

1. Dataset of real news was created from articles analyzed during three months[†] that were validated using https://mediabiasfactcheck.com. The dataset contained 15 707 articles.
2. Dataset of fake news was created[‡] based on the text analysis of 244 web pages marked as “bullshit” from BS Detector Chrome Extension 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[§] and processed based on [11]. The dataset was constructed using an end-to-end system, FakeNewsTracker [8]. Authors of this system collect the verified fake news and true news from fact-checking websites like PolitiFact on a daily basis. Then, using Twitter’s advanced search API, they gather the tweets related to the fake/real news that spread them on Twitter. This system served to create multi-dimensional information related to news content, social context, and spatiotemporal information. The output dataset from this system with detailed analysis is available as dataset KaiDMML. 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, it was taken as the best-created one and in analysis in this paper, it was used to verify the facts found in the first two datasets.

The processed datasets will be used to evaluate the statistically significant differences between the word categories representing the part of speech. In traditional grammar, a part of speech is a category of words (or, more generally, of lexical items) that have similar grammatical properties. Words that are assigned to the same part of speech generally display similar syntactic behavior, they play similar roles within the grammatical structure of sentences. Commonly listed English parts of speech are noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, and sometimes numeral, article, or determiner.

Morphological analysis was applied to the words to identify the morphological tags. This was done using the tool TreeTagger [12]. The TreeTagger is a tool for annotating text with part-of-speech and lemma information. The English Penn Treebank tagset was used with English corpora annotated by the TreeTagger tool [13–15], developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart [10]. The TreeTagger has been successfully used to tag German, English, French, Italian, Danish, Swedish, Norwegian, Dutch, Spanish, Bulgarian, Russian, Portuguese, Galician, Greek, Chinese, Swahili, Slovak, Slovenian, Latin, Estonian, Polish, Romanian, Czech, Coptic and old French texts and is adaptable to other languages if a lexicon and a manually tagged training corpus are available.

The English Penn Treebank tagset contains 36 morphological tags. The news from the examined dataset was assigned morphological tags using the TreeTagger. Some tags were not included in the experiment because they were tags with a small number of occurrences or tags without further importance for this research: SENT (Sentence-break punctuation), SYM (Symbol), LS (list marker). In this experiment were used 33 morphological tags. These

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

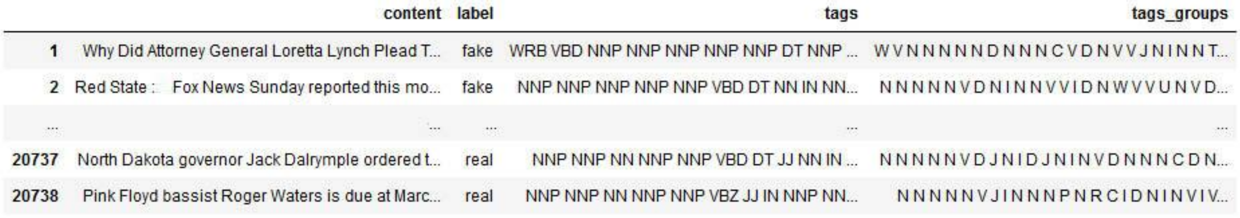
[‡] https://www.kaggle.com/mrisdal/fake-news

[§] https://github.com/KaiDMML/FakeNewsNet were then divided into 14 groups (e.g. into the group V were assigned all verbs, etc.). All created categories and all included tags are described in Table 1.

Table 1. Created groups of morphological tags

GTAG	POS Tags
group C	CC (coordinating conjunction) , CD (cardinal number)
group D	DT (determiner)
group E	EX (existential there)
group F	FW (foreign word)
group I	IN (preposition, subordinating conjunction)
group J	JJ (adjective) , JJR (adjective, comparative) , JJS (adjective, superlative)
group M	MD (modal)
group N	NN (noun, singular or mass) , NNS (noun plural) , NP (proper noun, singular) , NPS (proper noun, plural)
group P	PDT (predeterminer) , POS (possessive ending) , PP (personal pronoun)
group R	RB (adverb) , RBR (adverb, comparative) , RBS (adverb, superlative) , RP (particle)
group T	TO (infinitive ‘to’)
group U	UH (interjection)
group V	VB (verb be, base form) , VBD (verb be, past tense) , VBG (verb be, gerund/present participle) , VBN (verb be, past participle) , VBP (verb be, sing. present, non-3d) , VBZ (verb be, 3rd person sing. present)
group W	WDT (wh-determiner) , WP (wh-pronoun) , WP$ (possessive wh-pronoun) , WRB (wh-abverb)

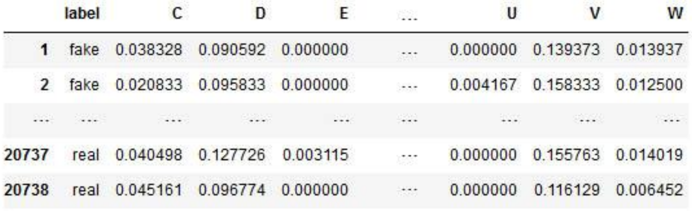
Fig. 1 depicts a sample of the dataset after the morphological tags identification and following morphological group’s extraction (part of speech).



	content	label	tags	tags_groups
1	Why Did Attorney General Loretta Lynch Plead T...	fake	WRB VBD NNP NNP NNP NNP NNP DT NNP ...	W V N N N N N D N N N C V D N V V J N I N N T...
2	Red State : Fox News Sunday reported this mo...	fake	NNP NNP NNP NNP NNP VBD DT NN IN NN...	N N N N N V D N I N N V V I D N W V V U N V D...
...	...	...	...	...
20737	North Dakota governor Jack Dalrymple ordered t...	real	NNP NNP NN NNP NNP VBD DT JJ NN IN ...	N N N N N V D J N I D J N I N V D N N N C D N...
20738	Pink Floyd bassist Roger Waters is due at Marc...	real	NNP NNP NN NNP NNP VBZ JJ IN NNP NN...	N N N N N V J I N N N P N R C I D N I N V I V...

Fig. 1 Dataset after the morphological tag’s identification

In the last step of data preparation, the relative counts of occurrence for each morphological tag was calculated for each news article from the examined dataset. Also, the relative counts of occurrence of each morphological group were assigned to the news articles of the examined dataset (Fig. 2).



	label	C	D	E	...	U	V	W
1	fake	0.038328	0.090592	0.000000	...	0.000000	0.139373	0.013937
2	fake	0.020833	0.095833	0.000000	...	0.004167	0.158333	0.012500
...	...	...	...	...	...	...	...	...
20737	real	0.040498	0.127726	0.003115	...	0.000000	0.155763	0.014019
20738	real	0.045161	0.096774	0.000000	...	0.000000	0.116129	0.006452

Fig. 2 A sample of the pre-processed dataset ready for analysis

4. Results

The aim of the research was to find out which word categories are typical for fake news and which for real news. It is to find out whether there are statistically significant differences in relative counts between word groups in the analyzed articles.

An analysis of variance was used for testing the differences among combinations of within-group and between-group factors (GTAG * fake) in relative frequencies of tags occurrence. The null hypothesis states that the relative frequencies of tags occurrence do not depend on the combination of GTAG and fake factors.

Table 2. Mauchly's sphericity test a) All Datasets b) KaiDMML

All Datasets	**W**	**Chi-Sqr.**	**df**	**p**
GTAG	0.00	702949.3	90	0.0000
KaiDMML	**W**	**Chi-Sqr.**	**df**	**p**
GTAG	0.00	10110.7	90	0.0000

The repeated-measures ANOVA was applied to data obtained by tagging (by the way of morphological analysis). Assumption of the analysis of variance for repeated measures is an equality of the variances and covariances in the covariant matrix for repeated measures, so-called assumption of the covariant matrix sphericity. The assumption of normality had not to be tested as the research sample was big enough.

To test the equality of the variances and covariances in the covariant matrix, Mauchly’s sphericity test was used (Table 2). The test is statistically significant ($p < 0.05$), the assumption of the equality of the variances is violated (Table 2). If the assumption of the covariance matrix sphericity is not fulfilled, the value of the type I error increases. In such cases, when applying F-test, the degrees of freedom are adjusted, with define corrections, to achieve the declared significance level. Due to the violation of the validity of the assumption of the variance analysis, Greenhouse-Geisser and Huynh-Feldt correction for analysis of variance repeated measures were used (Table 3).

Table 3. Adjusted univariate tests for repeated measure a) All Datasets b) KaiDMML

All Datasets	**df**	**F**	**p**	**G-G Epsilon**	**G-G Adj.df1**	**G-G Adj.df2**	**G-G Adj.p**	**H-F Epsilon**	**H-F Adj.df1**	**H-F Adj.df2**	**H-F Adj.p**
GTAG*fake	13	476.7	0.0000	0.1	1.9	55837.5	0.0000	0.1	1.9	55843.1	0.0000
KaiDMML	**df**	**F**	**p**	**G-G Epsilon**	**G-G Adj.df1**	**G-G Adj.df2**	**G-G Adj.p**	**H-F Epsilon**	**H-F Adj.df1**	**H-F Adj.df2**	**H-F Adj.p**
GTAG*fake	13	11.2	0.0000	0.2	2.4	959.4	0.0000	0.2	2.4	967.9	0.0000

Based on repeated measures ANOVA results (Table 3), we reject the null hypotheses at the 0.1% significant level, i.e. it was proven a statistically significant difference in relative frequencies of tags occurrence between combination of the GTAG factor and fake factor over all datasets (Table 3a) as well as for validation dataset KaiDMML (Table 3b).

Table 4. Multiple comparisons: All Datasets

fake	GTAG	Mean	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	U	0.0001	***																			
1	U	0.0002	***	***																		
0	F	0.0005	***	***	***																	
1	E	0.0016		***	***																	
0	E	0.0017			***																	
1	F	0.0032				***																
0	M	0.0103					***															
1	M	0.0107					***															
1	W	0.0137						***														
0	W	0.0151							***													
1	T	0.0249								***												
0	T	0.0261								***												
1	C	0.0481									***											
0	C	0.0490									***											
0	P	0.0736										***										
1	P	0.0744										***	***									
0	J	0.0756											***									
1	R	0.0775												***								
1	J	0.0787												***								
0	R	0.0788												***								
1	D	0.0892													***							
0	D	0.0954														***						
1	I	0.1116															***					
0	I	0.1205																***				
1	V	0.1574																	***			
0	V	0.1697																		***		
0	N	0.3426																			***	
1	N	0.3681																				***

After rejecting the global hypothesis, we are interested in which combinations of within-group factor and between-group factor (GTAG * fake) are statistically significant differences. The smallest occurrence of tags was found in groups U, F, and E (<1%) and the highest was found in group I, V, and N (> 10%) over all datasets (Table 4) as well as in the validation dataset KaiDMML (Table 5). The highest differences between the results over all datasets and the results on the validation dataset KaiDMML were found by the identification of statistically significant differences between fake and real news. In the case of all datasets (Table 4), we identified statistically significant differences in the tags´ occurrence between fake and real news for groups F (foreign words), W (wh-words), J (adjectives), D (determiners), I (prepositions), V (verbs), and N (nouns). In the case of the KaiD-MML dataset (Table 5), these results were confirmed only for groups V (verbs) and N (nouns), besides the difference was also identified in the case of group R (adverbs).

Table 5. Multiple comparisons: KaiDMML

fake	GTAG	Mean	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	U	0.0002	***													
1	F	0.0002	***													
0	F	0.0003	***													
0	U	0.0003	***													
0	E	0.0015	***	***												
1	E	0.0018	***	***												
0	M	0.0106		***	***											
1	M	0.0127			***											
0	W	0.0162			***	***										
1	W	0.0166			***	***										
0	T	0.0255				***	***									
1	T	0.0287					***									
1	C	0.0428						***								
0	C	0.0437						***								
1	J	0.0699							***							
0	J	0.0713							***							
0	P	0.0838								***						
0	R	0.0852								***						
1	P	0.0908								***	***					
0	D	0.0912								***	***					
1	D	0.0921								***	***					
1	R	0.0953									***					
0	I	0.1129										***				
1	I	0.1139										***				
0	V	0.1733											***			
1	V	0.1840												***		
1	N	0.3237													***	
0	N	0.3486														***

5. Discussion

This paper dealt with the fake and real news analysis that were obtained from three existing datasets of news articles. The used datasets were all open-source but originated from the year 2016. It is possible that the writing of fake news is evolving and the results of the analysis based on the used dataset do not have to take into account the newest trends of fake news. Despite that, the used dataset offers reliable articles that are suitable for various analyses.

Each word was assigned morphological tag and these tags were thoughtfully analyzed in this paper. The first step consisted of creating groups that consisted of related morphological tags. The groups reflected on the basic word classes.

The most important finding of this paper is the identified statistically significant differences in the use of word classes. Significant differences were identified (Table 4) for groups F (foreign words), J (adjectives) and N (nouns) in favor of fake news and groups W (wh-words), D (determiners), I (prepositions), V (verbs) in favor of real news. In the case of the third dataset that was evaluated separately and was used for verification, were identified significant differences for groups R (adverb), V (verbs), N (nouns).

When comparing the results of the examined dataset and the control dataset, it can be seen that statistically significant differences are identified for the word classes of verbs and nouns. Despite that the differences were identified only for some groups, significant differences can be identified if each tag is analyzed.

6. Conclusion

This paper analyzed the morphological tags and compared the differences in their use in fake news and real news articles. Statistically significant differences were identified for some tags in one or the other group. Despite that now it is not important for which morphological tags were identified as significant differences. It is important that the differences exist and it is obvious that morphological tags can be used as input into the fake news classifiers. Whether the relative counts are used as input layer of neural network or it is implemented into training data for other classifiers, the important finding is that this information can help to improve the classifier.

It is also important to note that morphological annotation is the basic (and most used) linguistic information put into corpora especially in the case of flective languages. The future work is oriented to the creation of its own dataset focused on articles in the Slovak language (the language of flective type). The dataset will be created using a web crawler that will browse the news reports websites daily, including websites marked as conspiracy ones, i.e. with increased probability of fake news articles occurrence. Selected articles will be evaluated by human, following a morphological classification and subsequently replicate the morphological tags analysis on the flective language. The aim would be to improve the existing web services for fake news classification. Finding statistically significant differences in individual categories of word classes is an important piece of information for the future fake news classifier in terms of selecting appropriate variables for the classification.

Acknowledgements

This work 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 under contract 313011T527.

References

[1] Niklewicz, K. (2017) Weeding Out Fake News: An Approach to Social Media Regulation. *European View*. 16 (2), 335–335.

[2] Xu, K., Wang, F., Wang, H., and Yang, B. (2018) A First Step Towards Combating Fake News over Online Social Media. in: Springer, Cham, pp. 521–531.

[3] Brașoveanu, A.M.P. and Andonie, R. (2019) Semantic Fake News Detection: A Machine Learning Perspective. in: Springer, Cham, pp. 656–667.

[4] Saikh, T., Anand, A., Ekbal, A., and Bhattacharyya, P. (2019) A Novel Approach Towards Fake News Detection: Deep Learning Augmented with Textual Entailment Features. in: Springer, Cham, pp. 345–358.

[5] Agudelo, G.E.R., Parra, O.J.S., and Velandia, J.B. (2018) Raising a Model for Fake News Detection Using Machine Learning in Python. in: Springer, Cham, pp. 596–604.

[6] Ahmed, H., Traore, I., and Saad, S. (2017) Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. in: Springer, Cham, pp. 127–138.

[7] Zhuk, D., Tretiakov, A., Gordeichuk, A., and Puchkovskaia, A. (2018) Methods to Identify Fake News in Social Media Using Artificial Intelligence Technologies. in: Springer, Cham, pp. 446–454.

[8] Shu, K., Mahudeswaran, D., and Liu, H. (2019) FakeNewsTracker: a tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory*. 25 (1), 60–71.

[9] Bhutani, B., Rastogi, N., Sehgal, P., and Purwar, A. (2019) Fake News Detection Using Sentiment Analysis. in: 2019 Twelfth Int. Conf. Contemp. Comput., IEEE, pp. 1–5.

[10] Dey, A., Rafi, R.Z., Hasan Parash, S., Arko, S.K., and Chakrabarty, A. (2018) Fake News Pattern Recognition using Linguistic Analysis. in: 2018 Jt. 7th Int. Conf. Informatics, Electron. Vis. 2018 2nd Int. Conf. Imaging, Vis. Pattern Recognit., IEEE, pp. 305–309.

[11] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2018) FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media.

[12] Schmid, H., Baroni, M., Zanchetta, E., and Stein, A. (2007) The Enriched TreeTagger System. in: Proc. EVALITA 2007 Work., .

[13] Munková, D., Munk, M., and Vozár, M. (2013) Data Pre-processing Evaluation for Text Mining: Transaction/Sequence Model. *Procedia Computer Science*. 18 1198–1207.

[14] Munková, D., Munk, M., and Adamová, L. (2014) Modelling of Language Processing Dependence on Morphological Features. in: V.

Trajkovik, M. Anastas (Eds.), ICT Innov. 2013, Springer International Publishing, Heidelbergpp. 77–86.

[15] Munkova, D., Stranovska, E., and Munk, M. (2015) Language Processing and Human Cognition. in: Huang, DS and Han, K (Ed.), Adv. Intell. Comput. Theor. Appl. ICIC 2015, PT III, pp. 500–509.

[16] Schmid, H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. in: Proc. Int. Conf. New Methods Lang. Process., Manchester, UKpp. 44–49.