Journal of Mathematics and System Science 9 (2019) 95-99
doi: 10.17265/2159-5291/2019.04.001

Official Statistics as Clickbait—The New Threat in the Post-truth Society?

Lyubomira Dimitrova

Department of Public Administration, Sofia Univeristy "St. Kliment Ohridski", 125 Tsarigradsko shose blvd, bl. 4, 413, 1113 Sofia, Bulgaria

Abstract: The aim of this paper is to raise awareness on the consequences of dissemination of official statistics through online media that uses clickbait headlines to generate traffic. In order to tackle on this issue, a Natural Language Processing (NLP) model was developed in order to distinguish the clickbait headline from the non-clickbait one when it comes to articles presenting information from the Bulgarian National Statistical Institute press releases. The yielded results are rather satisfactory as the parts-of-speech features model achieved an accuracy for 92% of the cases.

Key words: Official statistics, online media, clickbait.

1. Introduction

"Post-truth" has emerged as a popular term, referring to a particular way information has been presented to the public. According to the definition given by the Oxford dictionary, it refers to a situation in which objective facts are being set aside to more emotionally shaped information. Even though public bodies, such as the European Commission, are fighting this approach to information sharing there is still a level of ambiguity when it comes to defining which source of information is trustworthy and which is not [1]. Under this circumstance official statistics are considered to be the tool that can help tackle with fake and exaggerated information [2].

Even though usually official data are used to oppose unreliable information, it may be the case that fake news publishers use data produced by official statistics as a tool for gaining popularity. This occurs when online media, known for publishing fake and low quality information still shares the content of press releases of the official statistics. In such a situation, the body text remains unchained and valid; however, the headline that precedes it is structured as a clickbait—content, the main purpose of which is to drag attention and to generate more views [3]. Such practices are threatening to jeopardize the trust in official information presenting it in the same manner as the fake one.

An example can be seen with the media publications, following the release of the data from the labor cost survey of the Bulgarian NSI in 2018, which contains information on the average wage in the country. Instead of only sharing the published information, same headlines are going one step further framing the title as: "This is a must-read if you are working in Bulgaria! Very important information for the Bulgarian wages! The unthinkable has happened: Shocking". This case cannot be defined as fake news per se, since the content of the article is taken directly from the website of the National Statistics Institute of Bulgaria and the headline does not contain any data. Yet such articles tend to mislead the reader and discredit the validity of the data contained in the body text. Such approach to official information threatens to downgrade the level of trust in it and to blur the line between what is verified news and what is not.

Corresponding author: Lyubomira Dimitrova, PhD candidate, research field: text mining.

The aim of the paper is to develop a clickbait-detecting model, using data from all the headlines of articles containing press release information issued by the Bulgarian NSI from 21 media websites for 2017. Two models for clickbait detection are compared using different features—the first one uses words as features and the second one uses the method applied by Ref. [4] which uses type labels to frame the main features which a clickbait is containing, but for the purposes of the paper they have been converted into parts of speech. The reason why these approaches are chosen is that the former is considered easy to implement and simple, and the latter employs the most common features that a clickbait has—its dynamics, pathos and expression which can be detected by the parts of speech used. As the dataset is rather small and unbalanced in terms of share of clickbait vs. non-clickbait headlines (the former are fewer) an SVM (support vector machine) classifier was used. The results show the superiority of the parts of speech features, which is accurate in 92% of the cases, compared to the word feature model which predicted correctly 67% of the cases tested.

2. Experimental Section

This section contains information on the previous research on this topic highlighting the main findings in fake news and clickbait detection Natural Language Processing (NLP) models. The sub-section containing the literature review is followed by the data gathering process of this research and its design.

2.1 Literature Review

The topic of clickbait detection has attracted a number of scholars in the past several years and yet the term remains rather hard to define. Ref. [3] points out that due to the increased online media competition headlines are designed to be as appealing as possible to the audience in order to attract more viewers and to generate more traffic. In this context, Ref. [5] frames clickbait as a source of tabloidization of media, where otherwise valid information is presented in the same manner as the fake one. According to Ref. [6] the use of clickbait headlines leads to a decrease in the source of credibility for the reader. Thus even if a statistical authority produces accurate data, when published in such websites, it is expected to be treated in the same manner as the invalid and untrue information.

In terms of models used previously, researchers have developed various structures, focusing on features such as the length of words, their frequency, use of stop words, etc. Among the most common approaches, used by scholars such as Ref. [7] is the bag-of-words model, which traces the frequency of words used. It is easy to interpret and does not require complicated preprocessing of the data. Another approach, used by Ref. [4] is the transformation of sentences into sequence pattern in order to catch the most common features of ambiguous headlines.

When it comes to the building of an NLP model the language used plays a significant role. The majority of these models have been trained using articles and titles in English. So far there is only one fact checking model in Bulgarian, developed by Ref. [8] and it can be used to distinguish humorous from serious news. Thus when the text mining model has been developed for other language usually the dataset has to be built from a scratch. Ref. [9] has used the data gathered from a Hackaton in order to perform their fake news and click bait detection filter. For the development of their model they use four different types of features: lexical, stylometric, grammatical, embedings, following some of the most common clickbait characteristics. Their results yield 79.9% accuracy for the model using lexical features, hence it can be assumed that the term frequency–inverse document frequency (TDF IDF) score of a word feature can have a significant influence on the appearance of a clickbait article. However, as the authors confirm, as the TDF IDF trends are constantly changing the re-training of the model over time is necessary.

2.2 Methods Used

Compared to the model [8], this one has much more limited purpose—its task is not to detect clickbait titles in all their possible forms and contexts but to focus solely on those that had mentioned official data. Thus the dataset is collected from 21 news websites, covering all the headlines, preceding the publication from the Bulgarian NSI for 2017. Using the "rvest" library for web scraping in R, a total of 1,176 headlines were collected.

The next step for building the model was to distinguish between clickbait and non-clickbait titles. The criteria used were based on the findings of Ref. [10] regarding the main clickbait features. Some of them were not applicable for this case, as being highly inappropriate for announcing data from official statistics, the clickbait headlines were missing phrases such as "WOW" and "LMAO". In order to keep the model as simple as possible, three features of clickbait headlines were used for this research—the use of hyperbolic words, the punctuation patterns such as "!?", "...", "!!!" and the use of catchphrases such as "Unbelievable", "Shocking", etc. The result was the identification of 230 clickbait and 946 non-clickbait headlines.

The next task is to model the dataset according to the features used for the bag-of-words model, which follows the standard procedure for text preparation. Uppercases, numbers and stop words were removed. However, when the Parts-of-speech features model was prepared for some additional work needed to be done. Firstly, there was no need for removing the stop words as they were transformed into their parts of speech function and added into the corpus. Following the findings of Ref. [10], according to whom clickbait titles are characterized with higher levels of dynamics and usage of stop words, I converted the features into parts of speech, so that "The inflation has grown dramatically" becomes "Noun, verb, adverb". In addition, punctuation (except for comma, semicolon, colon and full stop) is also added into the pattern, labeled as "punctuation". Due to the poor variety of words in this case I expected to find a major difference in the use of verbs, adjectives, punctuation and interjections when it comes to the distinction between clickbait and non-clickbait titles.

When choosing the appropriate classifier, I took several specifics of the dataset into account. First of all, the overall number of headlines is rather small. Second, the majority of the features of the non-clickbait corpus occur in the clickbait one as well, as they announce the same press release. And last but not least, non-clickbait headlines are around four times as many as the clickbait ones, which make the training set very imbalanced. All this put together determined my choice of classifier for the model building, as the SVM can handle all these particularities. It is unbiased towards recurring features, works well with small datasets and thanks to the possibility to apply weights to the different classes it gives a solution to the imbalanced distribution of the training set [11]. In order to handle the unbalanced dataset, I added weights based on the reversed overall share of each of the subgroups in the dataset, using the class weight feature of the SVM function in R. The last step before the conduction of the model was to divide the dataset into training and test set, for both models that resulted into 940 observations for the training set and 236 observations for the test set.

3. Results and Discussion

The parts of speech model managed to predict correctly 92% of the cases from the test set, and the bag-of-words—only 67%. However, the accuracy rate in itself is not sufficient to measure the performance and that is why the two models are evaluated based on four other criteria. The precision shows the success rate for the prediction of positive observations from the total positive observations; recall is the ratio of correctly predicted positive observations to the all observations in actual class that is from all the clickbait (non-clickbait) articles how many were labelled correctly,

Table 1 Results from the models.

	Precision	Recall	F1	Area under the curve (AUC)
Words				
Clickbait	0.33	0.63	0.43	0.66
Non-clickbait	0.88	0.68	0.77	0.66
Parts of speech				
Clickbait	0.96	0.52	0.68	0.76
Non-clickbait	0.89	0.99	0.94	0.76

Source: author's calculation.

F1 is the ratio between the two and can be used as a criterion for accuracy. AUC stands for the area under the receiver operating characteristic curve, which plots two parameters—the true positive and the false positive rate and it shows the overall performance of the model. It ranges from 0 to 1, where 0 stands for a model unable to make any correct predictions and 1 is a model which makes only correct predictions. Usually AUC scores above 0.60 are considered as satisfactory.

Table 1 shows the overall results from the two models. What can be seen is that for both cases the models are performing better when predicting non-clickbait headlines. However, in this case the part of speech features model is performing much better, reaching a recall of 0.99 for the non-clickbait recognition. When it comes to the prediction of clickbait headlines, the performance can be considered as good, although not perfect, with F1 score of 0.43 for the words-as-features model and 0.68 for the parts of speech as features model. Looking at the AUC score of both models it can be concluded that the used parts of speech as features yield better results. As the results show the use of adjectives and punctuations in a headline it can solely be considered as a powerful feature for clickbait identification when it comes to official statistics presentation. It is easy to compute and overcome the usual language barrier for NLP models.

4. Conclusions

This paper aims at drawing attention to the hazardous effect low quality media has on official statistics through the dissemination of otherwise valid information as if it was fake. This approach is possible through the use of clickbait features in the headlines to attract more viewers thus presenting official statistical data and unconfirmed information in the same way. The communication channel for sharing statistical information is becoming biased and this may lead to a decrease in the level of trust the population has in the information from the official statistical sources. The second task of the research is to propose a possible solution for this problem of official statistics that is to develop a machine learning algorithm that can draw the line between a clickbait and a non-clickbait headline. The results show that clickbait headlines can be detected through the words used and their role as parts of the speech. Clickbait detection can help the users of official statistics to choose proper sources of information and to increase their level of trust in official data. It can also be used by the producers of data themselves as a media monitoring tool tracking those news channels that threaten to jeopardize their role as official information source.

Acknowledgment

I would like to thank Faculty of Philosophy of Sofia University for the financial support.

References

[1] Jaster, R., and Lanius, D. "What Is Fake News?" *Versus* 47 (2): 207-24.

[2] Baldacci, E. M. A. N. U. E. L. E., and Pelagalli, F. E. L. I. C. I. A. 2017. "Communication of Statistics in Post-truth Society: The Good, the Bad and the Ugly." *Publications Office of the European Union*, doi 10: 553958.

[3] Zheng, H.-T., et al. 2017. “Boost Clickbait Detection Based on User Behavior Analysis.” Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data. Springer, Cham. Science of Learning, National Science Teachers Association, Washington, D.C., pp. 85-7.

[4] Wei, W., and Wan, X. J. 2017. “Learning to Identify Ambiguous and Misleading News Headlines.” arXiv preprint arXiv:1705.06031.

[5] Chen, Y. M., Conroy, N. J., and Rubin, V. L. 2015. “Misleading Online Content: Recognizing Clickbait as False News.” In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*.

[6] Hurst, N. 2016. “To Clickbait or Not to Clickbait? An Examination of Clickbait Headline Effects on Source Credibility.” Thesis, University of Missouri.

[7] Potthast, M., et al. 2016. “Clickbait Detection.” European Conference on Information Retrieval. Springer, Cham.

[8] Hardalov, M., Koychev, I., and Nakov, P. 2016. “In Search of Credible News.” Presented at International Conference on Artificial Intelligence: Methodology, Systems, and Applications. Springer, Cham.

[9] Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., and Nakov, P. 2018. “Predicting Factuality of Reporting and Bias of News Media Sources.” arXiv preprint arXiv:1810.01765.

[10] Chakraborty, A., et al. 2016. “Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media.” Presented at 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

[11] Yaoyong, L., Bontcheva, K., and Cunningham, H. 2009. “Adapting SVM for Data Sparseness and Imbalance: A Case Study in Information Extraction.” *Natural Language Engineering* 15 (2): 241-71.