Official statistics as clickbait – the new threat in the post-truth society?

Lyubomira Dimitrova (l\_dimitrova@ff.uni-sofia.bg)[[1]](#footnote-1)

**Keywords:** official statistics, online media, clickbait

# 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. The role of official statistics in such circumstances is under threat, as it has been previously pointed by Baldacci et al. [1]. In their paper, official statistics is opposed to fake news, describing the possible relations between the two and the future actions that need to be taken.

Even though Baldacci et al. [1] present the two sources of information as competing, 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, main purpose of which is to drag attention and to generate more views [2]. Thus the release of the newest data from the Labour cost survey of the Bulgarian NSI, which contains information for the average wage in the country, for instance, is framed 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 practices of online media are harmful for the data dissemination as it threatens to jeopardise the trust in the official statistical sources. In order to prevent the harmful effects from these actions, the aim of the paper is to present 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 using the method applied by Wei et al. [3] 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 as easy to implement and simple, and the latter employs to 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) a support vector machine (SVM) 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.

# Methods

## 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. Zheng et al. [2] point 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, Chen et al. [4] frame clickbait as a source of tabloidization of media, where otherwise valid information is presented in the same manner as the fake one. According to Hurst [5] the use of clickbait headlines leads to a decrease in the source 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 previous researchers have developed various structures, focusing on features such as the length of words, their frequency, use of stopwords, etc. Among the most common approaches, used by scholars such as Potthast [6] 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 Wei et al. [2] is the transformation of sentences into sequence pattern in order to catch the most common features of ambiguous headlines.

## Methods used

As previously stated, this research is focused on the development of two models for clickbait detection - the first one is a bag-of-words model, and the second one is based on the word labelling of the parts of speech. The dataset was collected from 21 news websites, covering all the headlines, preceding the publication of a press release from the Bulgarian NSI for 2017. Using the rvest library for web scrapping in R, a total of 1175 headlines were collected.

The criteria used to distinguish clickbait from non-clickbait were based on the findings of Chakraborty et al. [7] regarding the main clickbait features. Some of them were not applicable for this case, as though 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 945 non-clickbait headlines.

The bag-of-words model required a standard procedure for text preparation: removing the uppercases, numbers and stop words. The second corpus, containing the type labels, needed further transformation. Following the findings of Chakraborty et al. [7], 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. Due to the poor variety of words in this case I expect 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 Support Vector Machine (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. [8]

# Results

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, 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 positives 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. Performance of the models**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Precision* | *Recall* | *F1* | *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 |

Table 1 shows the overall results from the two models. What can be seen is that for both cases the models are preforming better when predicting non-clickbait headlines. However, in this case the Parts 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 using parts of speech as features yield at better results.

# Conclusions

This paper aims at drawing attention on 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.

# References

1. Emanuele Baldacci and Felicia Pelagalli, Communication of statistics in a post-truth society: the good the bad and the ugly, (Published in the present special Statistical Working Paper issue), (2017)
2. Zheng, Hai-Tao, Xin Yao, Yong Jiang, Shu-Tao Xia, and Xi Xiao, Boost clickbait detection based on user behaviour analysis, In Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data Springer, Cham (2017), pp. 73-80
3. Wei Wei and Xiaojun Wan, Learning to identify ambiguous and misleading news headlines (2017), arXiv preprint arXiv:1705.06031
4. Chen Yimin, Niall J. Conroy, and Victoria L. Rubin, Misleading online content: Recognizing clickbait as false news, In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection (2015), pp. 15-19
5. Nathan, Hurst, To clickbait or not to clickbait? An examination of clickbait headline effects on source credibility, PhD diss., University of Missouri--Columbia, 2016
6. Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen, Clickbait detection, In European Conference on Information Retrieval (2016), pp. 810-817.
7. A. Chakraborty, B. Paranjape, S. Kakarla and N. Ganguly, Stop clickbait: Detecting and preventing clickbaits in online news media, In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on (2016), pp. 9-16
8. Y. Li, K. Bontcheva, H. Cunningham, Adapting SVM for Natural Language Learning: A Case Study Involving Information Extraction, Natural Language Engineering, 15(02), (2009), pp. 241-271
1. Sofia University, Department of Public Administration, Sofia, Bulgaria [↑](#footnote-ref-1)