Ivan OPIRSKYY[1], Yevhen SHTEFANIUK[2], Ihor IVANCHENKO[3], Ihor VOSCHYK[4]

Opiekun naukowy: Ivan OPIRSKYY[1]

ANALYSIS OF APPLICATION OF EXISTING FAKE NEWS RECOGNITION TECHNIQUES TO COUNTER INFORMATION PROPAGANDA

Summary: This article considers the features of information propaganda and approaches to combating it; the effectiveness of several well-known techniques for recognizing fake news; the analysis of possible efficiency of these techniques in the context of possibility of their application for counteraction to purposeful information influences was carried out. Based on the study, the most promising algorithm for recognizing information propaganda in social networks was selected.

Keywords: fake news, neural networks, information influence counteraction, information propaganda, Fakedetector, UDF, HC-CB-3.

ANALIZA STOSOWANIA ISTNIEJĄCYCH TECHNIK ROZPOZNAWANIA FAŁSZYWYCH WIADOMOŚCI DO PRZECIWDZIAŁANIA PROPAGANDZIE INFORMACYJNEJ

Streszczenie: W artykule omówiono cechy propagandy informacyjnej i sposoby jej zwalczania; skuteczność kilku dobrze znanych technik rozpoznawania fałszywych wiadomości. Dokonano analizy możliwej skuteczności tych technik w kontekście możliwości ich zastosowania do przeciwdziałania celowym wpływom informacyjnym. Na podstawie badania wybrano najbardziej obiecujący algorytm rozpoznawania propagandy informacyjnej w sieciach społecznościowych.

Słowa kluczowe: fałszywe wiadomości, sieci neuronowe, przeciwdziałanie wpływowi informacji, propaganda informacyjna, wykrywacz fałszywych wiadomości, UDF, HC-CB-3

[1] DSc, Lviv Polytechnic National University, Associated Professor of Department of Information Protection, iopirsky@gmail.com

[2] Lviv Polytechnic National University, Postgraduate student of Information Protection Department, yevhen.sht@gmail.com

[3] PhD, National Aviation University, IT-security Academaic Depatment, igor-p-l@ukr.net

[4] National Aviation University, IT-security Academaic Depatment, ivoschyk@gmail.com

1. Introduction

In recent years, we have seen a tendency to increase the role of social media in the lives of Internet users. More and more people are using social media and online resources to get the latest news. As a result, one of the main challenges is to determine the level of trust in the information spread by these resources.

Recently, so-called fake news - articles that spread false information, written specifically to mislead the user, - became more popular. According to some sources [1,2], in some cases the number of such fake news may exceed the number of true news, which creates the so-called effect of information vertigo, when users can no longer distinguish true information from fictional. This becomes the basis for a powerful informational impact on social opinion. A striking example is research that points to the direct impact of the spread of fake news on the US presidential campaign and the election of Donald Trump [3]. And if some fake news are simply designed to mislead a certain subject or event, the deliberate spreading of false information by certain organizations or governments can influence political decisions or even destabilize the situation in society.

According to [4], one of the main problems in the EU and the US in cyberspace is information propaganda from some post-Soviet countries. It poses a serious threat in both social and geopolitical contexts.

Thus, the development of effective means of counteracting such information propaganda, both by individual organizations and by governments, is extremely important.

The aim of the study was to investigate the features of information propaganda from the post-Soviet countries; propose an approach to counter such propaganda; to analyze existing algorithms for recognizing fake news and their effectiveness and choose the most accurate, from the point of view of the authors, an algorithm that would build an effective system for implementing the proposed approach to counter such information propaganda on social networks.

1.1. Main part

Information propaganda from some post-Soviet countries has a number of features that distinguish it from traditional information warfare. Researches conducted before allowed to define these features and also to present techniques applied by them.

Main features of information propaganda from post-Soviet countries include [5]:

- an unusually large amount;
- inconsistency;
- a large number of distribution channels;
- distortion of real facts, and sometimes their complete falsification.

Let's look at each of them.

The large amount of fake information allows it to be confirmed by a large number of users, which increases the level of trust in such information. In addition, according to recent studies, in the absence or little interest from end users, the degree of persuasiveness depends more on the number of facts that support the information than on the degree of veracity of each of them [5].

A large number of distribution channels provides opportunities to convey propaganda to users from different countries, from sources that are not related to the country of origin.

In addition, information obtained from various sources looks more truthful to the user. The inconsistency of propaganda is contrary to the classical notion of information warfare [5], where each message must be consistent with others and with the general ideology. However, the inconsistency of individual channels or individual messages of propaganda has its advantages: obtaining information from different points of view seems to increase trust in the source of information.

Distortion or falsification of real facts is one of the main factors that determine the effectiveness of such propaganda and the difficulty of counteracting it [5]. Since this approach does not require that fake messages contain any truthful information, the propaganda machine can react instantly to any world events, presenting them from a favorable side. And since the first message is the most impressive, subsequent messages can simply be skipped by the user. In addition, being the first, such a message is quickly spread by social media users, thus increasing the amount of fake information about the event.

The study was focused on social networks as channels for spreading false data, as they have a large audience, are able to quickly spread fake information and user reactions, and they can organize effective propaganda recognition using modern algorithms. Social networks can be used for several purposes - the spreading of fake information by bots or real users, discrediting existing trusted sources, escalating social tensions. To quickly detect such pages and fake information that they distribute, it is advisable to use algorithms for recognizing fake news, which will be discussed in the next section.

1.1.1. An overview of existing approaches to fake news recognition

Based on the characteristics of the data taken into account by the algorithms, they can be divided into:

- content-based: take into account textual information (ie the text of the news, a post on a social network or tweet, descriptions of profiles on a social network);
- social-based: take into account the social component (features of the distribution of the post on the social network and the reaction of users to it);
- combined: use both approaches..

Since social networks play a leading role in spreading, false information, the study is focused on the analysis of algorithms for detecting fake information in social networks.

An example of an approach that uses textual information from the author's post and profile is the FAKEDETECTOR framework [7]. It is based on the results of research that show that there is a strong correlation between the author of the news, its content and the topic. This framework uses a hybrid feature learning unit (HFLU), which is used to identify the external and internal features of a particular news item. Next, FAKEDETECTOR uses a deep diffusive neural network to further process the vectors of these features. Each of the entities - the author, topic and content of the news - is assigned its own trust rating.

This approach uses RNN (Recurrent Neural Network) to highlight external and internal features. The obtained feature vectors fall into the deep diffusive neural network, which, based on the input data, calculates the confidence level for each entity. The authors of this approach propose to train the neural network by the method of inverse error propagation, as the most effective method for this task.

Thus, FAKEDETECTOR is a framework with prior learning that determines the level of trust in the information content itself, its author and its topic.

Another approach that uses combining algorithms to process both textual information from posts and information about user reactions to these posts is described in [6]. This method takes into account the peculiarities of the spread of fake news on social networks. He suggests using two separate methods to detect fake content. The first method is based on the assertion that false news can be detected by the number and characteristics of user reactions to it. As an algorithm that takes into account the social features of fake distribution, the authors propose to use the approach "harmonic boolean label crowdsourcing (HC) on social signals", or HC-CB-3. However, it is effective only when this content has been distributed on the social network for some time, and a sufficient number of users have responded to it. The fake just created on the social network at the beginning of its existence may not collect a small number of comments or reactions, or they may be absent at all, which makes it impossible to use this method. In this case, the authors propose to use an algorithm for analyzing the content of posts to identify features that are characteristic of fake information content. A crucial element for this approach is the value of the threshold for selecting two algorithms, which determines at what number of user reactions to use the first, and at what - the second method.

Thus, a feature of this approach is the combined use of two methods that take into account both the specifics of news content and the peculiarities of its distribution on the social network.

An example of an approach that combines the use of both text and social components in one algorithm is the algorithm described in [8] - unsupervised framework, or UFD. It is based on taking into account the reaction of users to certain news. It is believed that when commenting on a post, the user thus expresses his opinion about it (if the information is true or not to his opinion), and therefore, this data can be used as a factor in recognizing a fake. According to this approach, all users are divided into two groups - trusted and untrusted, depending on the accuracy of their assessment of specific posts, i..e the number of matches between their assumptions in the reaction and the true assessment of the veracity of the news. Depending on the specific group to which the user belongs, when analyzing information content, his reaction is taken into account with a certain factor.

As the main algorithm, authors use Collapsed Gibbs Sampling, update rule of which is calculated based on the number of user reactions from trusted and untrusted groups. This allows the algorithm to adapt to changing application conditions, but imposes certain restrictions on the presence of reactions to messages.

1.1.2. Analysis of the peculiarities of information propaganda

The main task of this study was to analyze the features of information propaganda, in particular, the spreading of false information through social networks; analysis of modern approaches to detect fake news on social networks and their effectiveness;

determining the most accurate approach for the effective detection of false information disseminated on social networks in the framework of information propaganda.

The aim of the study was to determine the most accurate approach to detecting fake news on social networks, which can be applied to false information spread through information propaganda. There are a number of features that are characteristic of this information channel:

- fake news is spread on social networks in the form of user posts;
- users can specify links to news on a particular Internet resource;
- It should be noted that users can be both agents of information influence, who knowingly create and distribute fakes, and unconscious disseminators of false information.

The goals pursued by propaganda can be generally divided into two areas: the formation of opinion in society about a particular event or process - then the number of news spreads comes to the fore, and the escalation of tension in society - then controversy in user reactions to this news.

In general, the goals for active propaganda are well-known and popular socially significant issues, or recent events in the country.

Therefore, a comprehensive approach is needed to effectively detect such fakes. Let's form the criteria that it must meet:

- detect false news at the earliest stage, with a small number of user reactions, or their complete absence;
- provide a high detection rate;
- take into account both the features of the content and user reactions to posts;
- since information propaganda is a purposeful process, the approach should take into account possible changes in the paradigm of influence, as well as the fact that some users of the social network may be so-called bots or agents of information influence.

The study analyzed the effectiveness of the above approaches and algorithms for recognizing fake news. Their training was based on several popular fake content databases: BuzzFeed, PolitiFact and LIAR [9]. From the point of view of counteracting information propaganda, the BuzzFeed database is the most optimal. It contains a wide variety of political facts, the source of which can be not only politicians, parties or organizations, and this is the peculiarity of informational influence - fake news can be attributed to specific people and the collective concept of "experts" or no author . The HC-CB-3 and UFD algorithms will be compared according to the BuzzFeed database, and the FAKEDETECTOR algorithm will be compared according to PolitiFact.

The following metrics were used [10]:

- accuracy - the ratio of the number of correct predictions to the total number of predictions;
- precision - the ratio of correctly predicted positive observations to the total number of expected positive observations;
- sensitivity (recall) - the ratio of correctly predicted positive observations to all observations in the real class
- criterion F1 - a weighted average between accuracy and sensitivity.

The authors of these algorithms provided data on the effectiveness of other popular algorithms and frameworks for comparison, in particular:

- TRIFN [11]: uses user concepts, news and the public to identify fake news;
- DEEPWALK [12] - network embedding model. Based on the structure of the fake news network, DEEPWALK embeds articles, their authors and headlines in the latent feature space;
- RST [13] - is based only on content and uses SVM classifier;
- LIWC [14] - is based only on content and uses psycholinguistic categories;
- Castillo [15] - based only on social characteristics and takes into account user profiles and a network of friends;
- RST + Castillo [15] - combines approaches of content and social characteristics, however, due to the relatively low efficiency, they were not considered in the study.

To compare the effectiveness of the three algorithms, necessary data from the studies of Fakedetector, HC-CB-3 and UFD were isolated. In particular:

- for the Fakedetector algorithm, the results of classification by content category were taken, because in order to counteract information propaganda, it is most important to first determine the veracity of the news content;
- for the HC-CB-3 algorithm, the final results were the average value of efficiency for groups of knowingly true and knowingly fake newsgroups.

The results of the efficiency of the Fakedetector, HC-CB-3 and UFD algorithms are shown in Fig. 1.

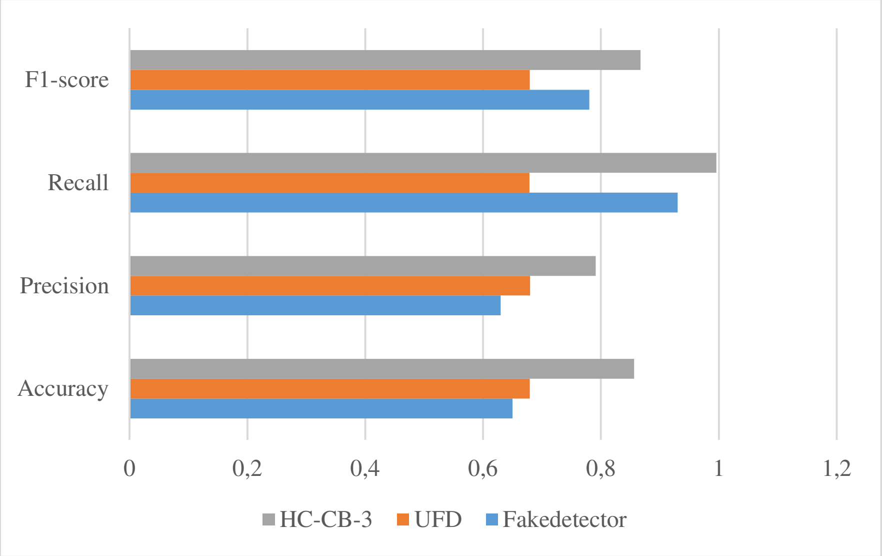

Figure 1. Efficiency results of Fakedetector, HC-CB-3 and UFD algorithms

As you can see, the most effective algorithm is HC-CB-3. It outperforms the UFD and Fakedetector algorithms according to all metrics used. In addition, HC-CB-3 has been implemented by the authors in the form of a chatbot for the social network Facebook and has already been tested on real data.

Thus, an approach with a comprehensive combination of content and social interaction can be effective for identifying information propaganda on social networks. It meets all the criteria we defined above: provides detection of false information in posts in the early stages, has a high detection rate of fakes, and takes into account both the content of the post content and social reactions of users to it. However, it requires a training sample for training, so we believe that this approach can be improved to work in an informational environment. To do this, we propose to dynamically collect data on the authors of already analyzed posts and store them in a separate database. This will allow you to build a system of evaluation of the author of the news and in the future to add to the algorithm the ability to take into account this assessment in the formation of the final decision. This improvement could make HC-CB-3 more effective in terms of outreach, as a lot of fake news is spread by specially created bot accounts.

2. Conclusion

In the course of the research of information propaganda peculiarities from some countries of the post-Soviet space were analyzed; modern approaches to the recognition of fake news in social networks and their effectiveness are considered; the most promising approach for effective detection of false information disseminated in social networks in the framework of information propaganda, which is the combined use of two algorithms that take into account different aspects of information content and its distribution in social networks - HC-CB-3. Its efficiency on the basis of fake news database BuzzFeed is higher than the efficiency of the considered algorithms UFD and Fakedetector in terms of accuracy, precision, recall and F1. However, we believe that the operation of the algorithm in the context of information propaganda can be improved if you create a database with estimates of trust in the authors of the posts and take it into account when deciding on new posts.

REFERENCES:

1. KATSAROS D., STAVROPOULOS G., PAPAKOSTAS D.: Which machine learning paradigm for fake news detection?, 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Thessaloniki, Greece, 2019, 383-387.
2. BRAILOVSKY M.M., IVANCHENKO I.S., OPIRSKY I.R., KHOROSHKO V.O.: Information and psychological confrontation in Ukraine, NAU: Scientific Journal "Information Security", 25(2019)3, Kyiv 2019, 144-149.
3. ALLCOTT H., GENTZKOW M.: Social media and fake news in the 2016 election. Journal of Economic Perspectives, 2017.
4. BOFFEY D.: EU raises funds to fight 'disinformation war' with Russia. The Guardian, 5 Dec 2018.
5. CHRISTOPHER P., MATTHEWS M.: The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It, Santa Monica, Calif.: RAND Corporation, PE-198-OSD, 2016. As of September 29, 2020: *https://www.rand.org/pubs/perspectives/PE198.html*
6. DELLA VEDOVA M. L., TACCHINI E., MORET S., BALLARIN G., DiPIERRO M, de ALFARO L.: Automatic Online Fake News Detection Combining Content and Social Signals, 2018 22nd Confer-ence of Open

Innovations Association (FRUCT), Jyvaskyla, 2018, 272-279, doi: 10.23919/FRUCT.2018.8468301.

7. ZHANG J., DONG B., YU P. S.: Fakedetector: Effective Fake News Detection with Deep Diffusive Neural Network, 2020 IEEE 36th International Conference on Data Engineering (ICDE), Dallas, TX, USA, 2020, 1826-1829, doi: 10.1109/ICDE48307.2020.00180.
8. YANG SHUO, SHU KAI, WANG SUHANG, GU RENJIE, WU FAN, LIU, Huan: Un-supervised Fake News Detection on Social Media: A Generative Approach. Proceedings of the AAAI Conference on Artificial Intelligence. 33(2019). 5644-5651. 10.1609/aaai.v33i01.33015644.
9. YINGZHAO OUYANG: Identifying fake news: The LIAR dataset and its limitations. Towards Data Science, June 29, 2020.
10. BROWNLEE J.: How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification - Machine Learning Mastery, January 3, 2020.
11. SHU K., WANG S., LIU H.: Exploiting tri-relationship for fake news detection. CoRR, abs/1712.07709, 2017.
12. PEROZZI B., AL-RFOU R., SKIENA S.: Deepwalk: Online learning of social representations. In KDD, 2014.
13. RUBIN V. L., CONROY N. J., YIMIN CHEN: Towards News Verification: Deception Detection Methods for News Discourse, 2015.
14. PENNEBAKER J. W., BOYD R. L., JORDAN K., BLACKBURN K.: The Development and Psycho-metric Properties of LIWC2015, 2015.
15. Castillo C., Mendoza M., Poblete B.: Information credibility on twitter, Proceedings of the 20th International Conference on World Wide Web. ACM, 2011, 675–684.