A Method of Recognition Technology of Fake News

Liudmyla Mishchenko, Valerii Simonenko
PhD Student, Ukraine, liudamishchenko@gmail.com
Professor, Associate Professor of Engineering, Ukraine

***Abstract* - the proposed article describes the method of recognizing fake news using Natural Language Processing and the Levenshtein Distance Algorithm technologies. It is proved that this method is effective for the task of recognizing fake or true news. Testing and comparative analysis results of the proposed method are given. It is theoretically and experimentally confirmed that the proposed method provides a good recognition of fake news level.**

***Index terms* - Natural Language Processing, fake news, the Levenshtein distance, tokens.**

I. INTRODUCTION

In today's world, television and news channels are increasingly sidelined as news and information are disseminated through the Internet and social networks. Users become subconscious victims of misinformation and manipulation.

In an information war, it is extremely important to critically apprehend and analyze information. But with the high pace of life and constant use of gadgets, users are reluctant to spend time repeatedly checking the facts. Therefore, using a mobile application or web browser extension is very relevant today and can greatly facilitate the fact-checking process.

To develop such a software product, it is necessary to use modern technologies that will help to quickly analyze and refute the facts. One of the best options for text analysis is to automatically synthesize natural language.

Natural language processing is used in all areas of modern technology. In most technological solutions, the recognition and processing of "human" languages has long been implemented: this is why a conventional IVR with rigidly answered answer options is gradually fading away, chatbots are starting to communicate more adequately without the involvement of a live operator, mail filters work on the basis of automated processes without human assistance etc.

Natural Language Processing (hereinafter referred to as NLP) is a natural language processing unit of computer science and AI dedicated to computer analysis of natural (human) language. NLP allows you to apply machine learning algorithms for text and language.

NLP can be used to create systems such as speech recognition, document generalization, spam detection, machine translation, question answering, named entity recognition, autocomplete, predictive text input, etc.

Today, almost 95% of smartphones have a tongue-in-cheek feature - they use NLP to understand human language. Also, most gadgets, such as laptops, tablets, are used with built-in language recognition.

The purpose of the work is to recognize the recorded language, that is, the text, using modern technologies. Automatic processing and analysis of collected data for fake or true facts or manipulation.

The use of modern technologies is a necessary factor in the fight against the spread of fake data. Moreover, the main task is the rapid automatic analysis of information, as well as the dissemination of denials and true facts. Therefore, developing new algorithms for searching and analyzing the news flow is an urgent task.

II. PROPOSED METHOD

From the above material, it follows that the use of Natural Language Processing technology to implement a way of recognizing fake news is extremely appropriate today. After all, this technology has not been used before for the purpose of combating misinformation.

The main objective of the article is to develop a way to recognize fake news. That is, the proposed approach to solving the problem should ensure that the true information is distinguished from the fabricated one. The solution should be able to proceed with a huge amount of facts.

Below is a summary of the algorithm for implementing the proposed method:

- Collection of verified and confirmed facts;
- Organizing the collected data in the database;
- Building list of meaningful tokens from texts and titles of true articles;
- Dividing the text of this article into words and forming a set of key tokens using NLP technology;
- Comparison of received tokens from the title of this text with tokens of titles of proven news by Levenstein's algorithm;
- Comparing article tokens that have a small difference in Levenstein distance in the title;
- Checking the results.

The European Union and the United States of America databases are used to collect information and create a data file. These databases contain links to verified sources and facts, as well as verified or denied news.

In order to optimize the data set, the collected data is cleaned from duplicates. In fact, when parsing different sources, there are cases of checking the same news several times, which increases the memory requirements and significantly slows down the system.

NLP technology is used to analyze the text, which forms a complete meaningful picture of the text. It splits the text for Universal Part-of-Speech Tagset. All of them are shown in Table 1 [1].

TABLE 1
UNIVERSAL PART-OF-SPEECH TAGSET

Tag	Meaning
ADJ	adjective
ADP	adposition
ADV	adverb
CONJ	conjunction
DET	determiner, article
NOUN	noun
NUM	numeral
PRT	particle
PRON	pronoun
VERB	verb
.	Punctuation marks
X	other

The title and text of the article are analyzed separately. The title provides a list of semantic tokens based on all words. Lists of semantic and key tokens are separately generated for the text. In doing so, the analysis removes from the text words such as determiner, article, particle, punctuation marks, other.

The process of forming semantic text tokens is shown in Fig. 1.

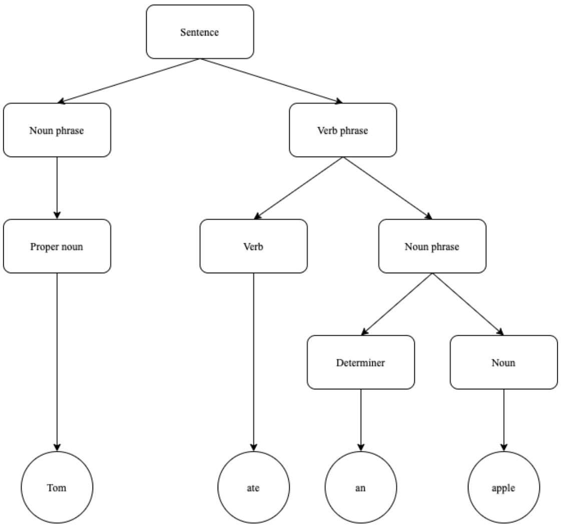

Fig 1. Semantic tokens parse tree

After the first analysis, there is one more action required before proceeding to comparing results. It requires removing redundant tokens from the text of the fact. These tokens usually do not contain any sensitive information, so this removal speeds up the whole execution of the algorithm. The final result is shown in Fig 2.

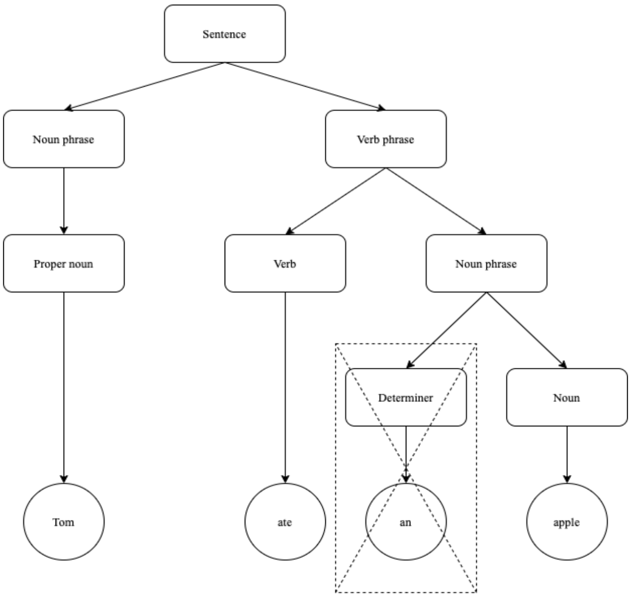

Fig 2. Semantic tokens parse tree after removal of redundant tokens

As you can see from the figure, NLP technology decomposes sentences into tokens. After that, not all meaning tokens are rejected and are not taken for further analysis. In this example, the determiner "an" was deleted, which does not make any sense.

After forming a list of tokens from the text and heading, a search for proven or simplified facts in the database is performed using the Levenstein algorithm. The Levenshtein distance is a number, used to measure difference between two character sequences, commonly used in information theory, computer science and linguistics. In general, it is the number of single-character edits (insertions, deletions or substitutions) which lead to transformation of one word to another. It was firstly introduced by Vladimir Levenshtein in 1965, and therefore got his name.

Levenshtein distance may also be referred to as edit distance, although that term may also denote a larger family of distance metrics known collectively as edit distance. It is closely related to pairwise all string alignments [2].

The Levenshtein algorithm calculates the least number of edit operations that are necessary to modify one string to obtain another one string. The most common way of calculating this is by the dynamic programming approach:

- A matrix is initialized measuring in the (*m*, *n*) cell the Levenshtein distance between the *m*-character prefix of one with the *n*-prefix of the other word.
- The matrix can be filled from the upper left to the lower right corner.
- Each jump horizontally or vertically corresponds to an insert or a delete, respectively.

- The cost is normally set to 1 for each of the operations.
- The diagonal jump can cost either one, if the two characters in the row and column do not match else 0, if they match. Each cell always minimizes the cost locally.
- This way the number in the lower right corner is the Levenshtein distance between both words [3].

Search of facts in the database goes in the following way:

- Facts are filtered by heading tokens, only facts which resemble header tokens greater than 75% according to Levenstein distance are chosen for next step;
- Split facts into similar and opposite by comparison of text tokens. When tokens are more than 50% similar, the fact is marked as similar. Otherwise, it is opposite.

The result of the actions taken is a list of true and simplified fakes, sorted by the relevancy to a given token list.

After that, taking the list of facts from db, fact is decided to be truth or fake depending on the following conditions:

- More than 90% of the facts of the given facts are similar and true. In this case, the fact is marked as “true” and recorded in the database;
- More than 90% of the facts of the given facts are similar and fake. In this case, the fact is noted as a "fake" and recorded in the database;
- More than 90% of the facts of the given facts are opposite and true. In this case, the fact is noted as a "fake" and recorded in the database;
- More than 90% of the facts of the given facts are opposite and fake. In this case, the fact is marked as “true” and recorded in the database;
- In all other cases, the fact is sent for manual verification.

III. EXPERIMENTAL DATA AND RESULTS

To test the given solution, news from the sputniknews site group were selected in dates range between February 1st, 2020 and February 29, 2020. The data from the euvsdisinfo.eu site was selected in order to build a database with approved facts. At the end of the scraping, the database of approved facts contained 7837 facts from different sources, including sputniknews site group, which was selected as target. 500 different news items were selected from this site group to be analyzed.

In average, it took about 2 seconds per each fact to run the algorithm. The results of the news analysis of the proposed method are shown in Table 2.

TABLE 2
THE RESULTS OF WORK OF THE PROPOSED METHOD

Category	Amount
Fakes	79
Manual verification	45
Truth	376
Total	500

As shown in the results table, 500 news items were analyzed. 79 of them turned out to be fake, 45 were sent for additional manual processing. The rest of the news was not found, so we believe they are true or in need of competent journalistic research.

After analyzing all chosen news based on the journalistic review, the results were as follows (shown in Table 3).

TABLE 3
RESULTS OF NEWS REVIEW BASED ON JOURNALISTIC REVIEW

Category	Amount
Fakes	95
Truth	405
Total	500

Thus, the program was able to correctly recognize 83% of fakes. Given the fact that the task of detecting fakes is crucial today, the results can be considered as positive and confirmatory expectations.

IV. CONCLUSIONS

This article describes how to spot fake news on the Internet. The method is different from the others using NLP technology and the Levenstein algorithm.

To analyze the proposed method, the database is filled with verified news from euvsdisinfo.eu. To test this method, a group of sputnicnews sites are selected, which usually spread fake news.

Tested a method that processed a sample of 500 news. Test results compared to the results of the fake news journalistic database. The method correctly recognized 83% of the fraudulent facts.

Thus, the proposed method proved to be effective for recognizing fake news with fairly high accuracy.

V. ACKNOWLEDGMENTS

The suggested way of recognizing fake news has shown rather high accuracy, but there are several ways to improve it.

For example, to recognize not only fakes, but also true news. And mark them as "true" for factual statements to users.

Also, this method can be used to create much faster and larger volumes of data than human resources alone. And then apply the generated datasets to train the neural network.

Fake in general also includes news types such as manipulation and partial truth. The difference between them can be explained as follows. In terms of intent, there are two main categories - disinformation and misinformation. The first is deliberate misrepresentation by the author (but certainly not by the disseminator), and the second by the unconscious (the user spreads false news). Partial truth can be both first and second. And manipulation is only the first. Practically, this can be converted to a manipulative index: if the source is reliable, then the lack of data is misinformation and, if not, disinformation. The media can also reflect not pure facts but their interpretation. Therefore, instead of reflecting reality, they create a pseudo-reality, marking and pinning the necessary information in it with the help of stereotypes. Therefore, recognizing all false information and classifying it in fakes, manipulation and partial truth is equally important. So, another way to improve is to recognize not only fakes but also manipulations and their separation.

VI. REFERENCES

Online book

[1] NLTK-book. Categorizing and Tagging Words: 2.3. A Universal Part-of-Speech Tagset [Online]. Available: http://www.nltk.org/book/ch05.html [Accessed January 16, 2020].

Articles in Online Encyclopedia

[2] The Levenshtein Algorithm [Abstract]. Available: medium.com, https://medium.com/cuelogic-technologies/the-levenshtein-algorithm-916db91843ba [Accessed February 26, 2020].

[3] The Levenshtein Algorithm. Dynamic Programming Approach [Abstract]. Available: cuelogic.com, https://www.cuelogic.com/blog/the-levenshtein-algorithm [Accessed February 28, 2020].