BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY
Excel@FIT 2024
http://excel.fit.vutbr.cz

Application for detection of Fake News

Bc. Jan Zádrapa

Abstract
The problem of Fake News is one of the most significant problems in our modern society. Millions of people read Fake News articles every day without knowing it. This problem creates a risk worldwide as society is getting polarised, and elections are manipulated by third parties using propaganda. Unfortunately, there are not enough tools to help solve the problem of Fake News detection in the Czech language. This thesis aims to create a tool for Fake News recognition of Czech articles with the help of natural language processing models.

*xzadra03@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Although most people know that Fake News is among us, many still need help distinguishing between serious and Fake News articles. The worst situation is between children and senior citizens [1]. Even serious news is getting angrier and more negative to attract attention, so the line between serious news and Fake News is getting thinner [2]. It is important to know what websites create these problems and which articles should be treated cautiously as the source is untrustworthy. This is the motivation for this work.

The main problem is defining what language techniques manipulate people the most or are the most common. These techniques should be detected with the help of machine learning, which means creating a suitable dataset for this topic. The goal is to create a tool for manipulation techniques detection in the input text (article) and tell if the article has signs of a Fake News article.

There is only one working solution in the Czech Republic called Verifee [3]. This web extension tool creates a score of the articles that have been read. The problem is, that this solution only works on a handful of websites.

The presented solution (FakeDet) uses the NLP pretrained model RoBERTa to detect selected manipulation techniques. It is implemented in a user-interactive version, so the user decides which text to evaluate. The result has a score from 0 to 100, like Verifee, where the score measures the trustworthiness of a website. It also shows the detected techniques in the input text. The application correctly recognises Czech Fake News articles, and even a web extension version shows that Fake News detection is possible even in a web environment.

2. Manipulation techniques

One way to recognise Fake News articles is through manipulation techniques. Serious news outlets should never use manipulation to gain readers. There are many manipulation techniques, but some are unsuitable for this type of work. There were selected four manipulation techniques, namely:

- fear,
- labelling,
- propaganda,
- conspiracy.

These techniques are selected as their severity is high or are common in the text. The technique of fear or appeal to fear is one of the most severe techniques because fear affects people's opinions and decisions. There are examples of elections affected by fear [4]. The second selected technique is labelling, commonly used in disinformation websites for people or countries. There are two types of labels, which are positive and negative. In the Czech Republic, the positive label usually has Russia or China; on the other hand, the negative one has USA or Ukraine [5].

The last two techniques are more complex, but their usage is common. Propaganda is a manipulation technique where the news is one-sided, meaning that one side is a hero in extreme cases, and the second is the villain. Conspiracy theories are a very well-known phenomenon that has its own specific language.

3. Natural language processing and dataset

When working with text, it is ideal to use natural language processing (NLP). There are many text tasks, like sentiment analysis, which NLP helps with. Sentiment analysis is the task of recognising some pattern in text and determining the positivity or negativity in text [6]. Sentiment analysis can also be used to detect manipulation techniques.

The work then uses and implements the NLP pipeline, a set of steps for achieving the quality NLP machine learning model, which is trained for specific tasks [7]. As seen in FIG 2, the NLP pipeline has several steps, including model training and data obtaining.

3.1 Dataset

Obtaining a dataset was difficult, as there are no publicly available datasets of labelled text for manipulation techniques detection. My consultant Mgr. Marek Grác, PhD, obtained a dataset from Masaryk University (MUNI) with several thousands of articles labelled with manipulation techniques [8]. Unfortunately, the dataset did not include all the four mentioned ones. This dataset comprises one-half of the 1035 data samples. The other half was scraped during the summer of 2023 from websites mentioned on the poster.

3.2 Model training

The dataset was then used to train NLP models for every manipulation technique. NLP pre-trained models BERT and RoBERTa [9, 10] were used as their proof of concept experiment on 100 data samples, with the best results. The average score is depicted in FIG 3 and shows almost no difference between these two NLP models for task-detecting manipulation techniques in text. All the used models in the application had at least an 80 % F1 score.

4. Implementation and testing

As visible in FIG 4, the main approach was to create an application where the user inputs text and the text is then evaluated on some scale. The scale is from 0 to 100 and shows the trustworthiness of an article.

There were two main versions of the tool, which is called FakeDet. One of the versions is a desktop application written in Python, where the user has to enter the input text manually. The second version is more focused on the automation of the process, and it is a Google Chrome extension that scrapes the text from the currently visited tab and evaluates it. The models in the background are the same. Precisely pretrained multilingual RoBERTa which is fine-tuned for this task. Figures FIG 5 and FIG 6 show the look of both versions of applications. Figure FIG 10 shows the categories to which the articles were divided.

The testing shows that the tool is working correctly. The table TAB 1 shows the scores of individual news pages. There is a visible difference between serious and Fake News sources. Also, there is a comparison with Verifee. Testing has shown that tools have a similar distribution of the scores. The distribution of the scores is depicted in figure FIG 7. The next test should confirm the correct behaviour, as there were two rounds of testing. The second round should confirm the consistency of the tool and be done with the whole article (almost, as there were even longer articles, but it was a minority of them). The testing results in two rounds are visible in figures 8 and 9.

5. Conclusion

The application works correctly, but there are considerations on how to upgrade the application. For example, the number of detected manipulation techniques could be higher, and the dataset could be bigger. However, the labelling process is time-consuming, so the 1000 data samples in the dataset are a success. Finally, the web extension only works with one technique and is not deployed on the server, so this could be a future challenge. Overall, this tool shows that it is possible to detect Fake News even in the Czech Republic.

Acknowledgements

I would like to thank my supervisor, Mgr. Kamil Malinka, Ph.D. for his support, advice, and helpful points while working on this thesis. I would also like to thank the consultant, Mgr. Marek Grác, Ph.D., for helping me in more technical matters and for getting me a dataset, which helped me very much.

References

[1] Brooke Adams. Aging and fake news: It's not the story you think it is. online, 2022.

[2] David Rozado, Ruth Hughes, and Jamin Halberstadt. Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with transformer language models. *PLOS ONE*, 17(10):1–14, 10 2022.

[3] Verifee. Verifee. online, 2023.

[4] Nikola Remešová. Migration topic in the pre-election programs of selected political parties/-movements of the parliament of the czech republic, 2019.

[5] Telewizja Polska S.A. Russian agent greased palms of czech public figures to spread russian propaganda. online, 2023.

[6] MonkeyLearn Inc. Sentiment analysis: A definitive guide. online, 2023.

[7] Pawankrgunjan. Natural language processing (nlp) pipeline. online, 2023.

[8] Vít Baisa, Ondřej Herman, and Aleš Horák. Benchmark dataset for propaganda detection in czech newspaper texts, 2019.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.