**MINISTERUL EDUCAŢIEI, CULTURII ȘI CERCETĂRII**

**Universitatea Tehnică a Moldovei**
**Facultatea Calculatoare Informatică şi Microelectronică**
**Departamentul Ingineria Software și Automatică**

**Admis la susţinere**
**Şef de departament:**
**Fiodorov I. dr., conf.univ.**

**-------------------------------**
„27” *octombrie* 2022

# METODE LINGVISTICE AUTO-SUPERVIZATE PENTRU DETECTAREA ARTICOLELOR FALSE DIN MASS-MEDIA

## Proiect de master

| | | |
|---|---|---|
| **Student:** | ____________ | **Speianu Dana, IS-211M** |
| **Coordonator:** | ____________ | **Gavrilița Mihail, asist. univ.** |
| **Consultant:** | ____________ | **Catruc Mariana, lect. univ.** |

**Chişinău, 2023**

## Rezumat

„Metode lingvistice auto-supervizate pentru detectarea articolelor false din mass-media” este tema cercetării pentru lucrarea de master realizată de către studenta Speianu Dana.

Cuvinte cheie: articole media false, procesarea limbajului natural, modele pre-antrenate de limbaj, fine-tuning, învățare prin transfer.

Această lucrare de cercetare a fost realizată pentru a studia eficiența utilizătii metodelor lingvistive auto-supervizate în sarcina de a detecta articolele media false. Astfel, s-au studiat metodele existente de detectare a articolelor media false, studiind și aspectele favorabile, cât și cele mai puțin favorabile. Utilitatea inteligenței artificiale crește cu pași majori în ultimul deceniu, fiind folosită pentru eficientizarea și îmbunătățirea proceselor ce anterior erau realizate manual de către oameni sau alte metode mai puțin eficiente. În domeniul procesării limbajului natural, deasemenea, au avut loc progrese majore mai ales în direcția tehnologiilor ce sunt folosite pentru reprezentarea cuvintelor ca vectori, deoarece calculatorul nu înțelege textul, pentru a transmite informațiile către modelul de clasificare. Există modele fără context care generează o reprezentare a unui singur cuvânt pentru fiecare din vocabular, dar sunt modelele contextuale (cum ar fi BERT, GPT) care iau în considerare și contextul în care este poziționat cuvântul. Aceste modele sunt pre-antrenate pe o cantitate imensă de date și generează vectori mai performanti pentru cuvinte. Marile companii ce au acces la o cantitate imensă de date au dezvoltat modele lingvistive ce au fost pre-antrenate pe acestea. Astfel, aceste modele pre-antrenate transformă cuvintele în vectori mult mai valoroși, prin urmare obținând performanțe remarcabile în sarcinile în care sunt aplicate.

În această lucrare au fost aplicate versiuni mai restrânse ale modelelor pre-antrenate originale, precum DeBERTa, GPT-2, GPT-J, YOSO, XLNet, RoBERTa, dar au fost obținute rezultate foarte bune. Modelele pre-antrenate au fost aplicate folosind două metode: extragerea caracteristicilor și fine-tuning(reglare minuțioasă). Utilizând metoda de fine-tuning au fost obținute rezultate excelente precum acuratețe de 99.87%.

Cercetările au arătat că utilizarea acestor modele pre-antrenate este o soluție bună pentru creșterea performanței. Multe aplicații au înregistrat progrese remarcabile ca urmare a dezvoltării și aplicării modelelor de limbaj pre-antrenate, la fel cum s-a realizat și în cadrul sarcinii din această lucrare. Acesta este încă un motiv pentru a considera că modelele lingvistice pre-antrenate sunt dintre cele mai bune metode de a obține performanțe ridicate pentru diferite sarcini.

# Abstract

„Self-supervised language models for detecting fake media articles” is the research topic for the master's thesis carried out by the student Speianu Dana.

Keywords: Fake media articles, Natural language Processing, Pre-Trained Language Models, fine-tuning, transfer learning.

This research paper was carried out to study the effectiveness of using self-supervised linguistic methods in the task of detecting fake media articles. Thus, the existence of detecting fake media articles was studied, studying both the favorable and the less favorable aspects. The utility of artificial intelligence is growing by leaps and bounds in the last decade, being used to streamline and increase processes that were previously done manually by humans or other less efficient ones. In the field of natural language processing, also, there have been major advances especially in the direction of technologies that are used to represent words as vectors, since the computer does not understand the text, to transmit the information to the classification model. There are context-free models that generate a representation of a single word for each word in the vocabulary, but it is the contextual models (such as BERT, GPT) that also take into account the context in which the word is positioned. These models are pre-trained on a huge amount of data and generate better performing word vectors. Big companies that have access to a huge amount of data have developed language models that have been pre-trained on it. Thus, these pre-trained models transform words into much more valuable vectors, thereby achieving outstanding performance in the tasks in which they are applied.

In this work, more restrained versions of the original pre-trained models, such as DeBERTa, GPT-2, GPT-J, YOSO, XLNet, RoBERTa, were applied, but very good results were obtained. Two methods were applied to the pre-regenerated models: feature extraction and fine-tuning. Using the fine-tuning method excellent results such as 99.87% accuracy were obtained.

Research has shown that using these pre-trained models is a good solution for increasing performance. Many applications have made remarkable progress as a result of the development and application of pre-trained language models, as was the case in this work. This is yet another reason to consider pre-trained language models to be among the best methods to achieve high performance for various tasks.

## Table of contents

# LIST OF FIGURES

# ABBREVIATIONS

AI - Artificial Intelligence

NLP - Natural Language Processing

NER - Named Entity Recognition

NLTK - Natural Language Toolkit

QA - Question Answering

PTM - Pre-Trained Model

GLoVE - Global Vectors for Word Representation

TF-IDF - Term Frequency — Inverse Document Frequency

BERT - Bidirectional Encoder Representations from Transformers

GPT - Generative Pre-trained Transformer

ELMo - Embeddings from Language Models

LSTM - Long Short-Term Memory

RoBERTa - Robustly Optimized BERT Pretraining Approach

YOSO - You Only Sample Once

# INTRODUCTION

Mass media always had a huge influence on people’s behavior, beliefs, and opinions. People constantly read a lot of articles, and news from a multitude of sources including magazines, and news websites. These articles promote attitudes, moods, and a sense of what is and are not significant. Fake news as a weapon or means of manipulation has always existed, but only in the 20th century, in the era of mass media propaganda, did it become a phenomenon of alarming proportions. As fake news spreads more frequently, people are becoming less likely to believe it when it is true, which diverts their attention from the real problem.

There are a lot of approaches for dealing with fake articles and the most challenging part is to detect if an article/news is fake or not. A way is to detect these articles manually by human detection, but this is not very accurate and takes a lot of time and effort. Another way is to detect these fake articles by using machine learning approaches. These solutions proved a great performance, all that is needed is a dataset with fake and true news, a good dataset pre-processing, a suitable word embedding model, and a classification model. The performance of the solution anyway depends on every step mentioned earlier. For this thesis, the main purpose is to apply recent language pre-trained models for more qualitative word embeddings. These are context-based, which means the models take into account also the context of the sentence too. These language models are trained on huge datasets and often are created by big corporates such as Google, Facebook, and Microsoft, which have access to a lot of data. This paper presents the research and analysis of the application of these language models for the classification of fake and true mass-media articles.

The paper has, for now, the first part:

**Domain analysis and research definition** - This chapter presents a detailed domain background, the thesis research scope, and objectives. Also, this chapter presents the analysis of existing solutions for detecting fake articles from mass media.

**Research process description** - This chapter represents the dataset pre-processing and exploration. The description of the modeling process by using different types of pre-trained models. There is done an analysis of two different approaches of pre-trained models usage in fake media articles detection. Besides this, the chapter presents the results and evaluation of the models.

# BIBLIOGRAPHY

1. Mitcham, D., Taylor, M., Harris, C. Utilizing Social Media for Information Dispersal during Local Disasters: The Communication Hub Framework for Local Emergency Management. Int J Environ Res Public Health. 2021 Oct 14;18(20):10784. doi: 10.3390/ijerph182010784. PMID: 34682529; PMCID: PMC8535717.

2. Ihsan, A., Mohamad, N. B. A., Palaiahnakote, S., Nurul, F. B. M. N., "Fake News Detection Techniques on Social Media: A Survey", Wireless Communications and Mobile Computing, vol. 2022, Article ID 6072084, 17 pages, 2022. https://doi.org/10.1155/2022/6072084

3. Medeiros, F. D. C., Braga, R. B., “Fake news detection in social media: a systematic review,” The ACM International Conference Proceeding Series, vol. 3, no. 5, pp. 2–7, 2020.

4. Meneses Silva, C. V., R. Silva Fontes, and Colaço Júnior M., “Intelligent fake news detection: a systematic mapping,” *Journal of Applied Security Research*, vol. 16, no. 2, pp. 168–189, 2021.

5. Qasim, R., W. H., Bangyal, Alqarni, M. A., and Ali Almazroi A., “A fine-tuned BERT-based transfer learning approach for text classification,” *Journal of healthcare engineering,* vol. 2022, Article ID 3498123, 17 pages, 2022.

6. Brownlee J., “What Are Word Embeddings for Text?” on October 11, 2017, Deep Learning for Natural Language Processing, last updated on August 7, 2019.

7. Goodfellow, I., Bengio, Y., Courville, A., Deep Learning (Adaptive Computation and Machine Learning series), page 538, 2016

8. Martinez, D., “Is Transfer Learning the final step for enabling AI in Aviation?”, April 4th, 2020. Available: https://datascience.aero/transfer-learning-aviation/

9. Li, H., “Language Models: Past, Present, and Future” *Communications of the ACM*, July 2022, Vol. 65 No. 7, Pages 56-63, 10.1145/3490443.

10. Yalçın, O. G., “3 Pre-Trained Model Series to Use for NLP with Transfer Learning”, December 5th, 2020.

11. "Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing". Google AI Blog.(Available: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html)

12. Rani, H., “BERT Explained: State of the art language model for NLP”, Towards Data Science, 2018.

13. Devlin, J., Chang, M.-W., K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *Google AI Language*, 24th July 2019.

14. Devlin, J., Chang, M.-W., “Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing”, *Google AI Language*, November 2nd, 2018. (Available: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html)

15. Chen, W., Zhang, Y., Yeo, C.K., Lau, C.T., Sung Lee, B. (2018) Unsupervised rumor detection based on users’ behaviors using neural networks. Pattern Recogn Lett 105:226–233.

16. Yang, F., Liu, Y., Xiaohui, Y., Yang, M. (2012) Automatic detection of rumor on Sina Weibo. In: Proceedings of the ACM SIGKDD workshop on mining data semantics, pp 1–7.

17. Roy, A., Basak, K., Ekbal, A., Bhattacharyya, P. (2018) A deep ensemble framework for fake news detection and classification. arXiv:arXiv-1811.

18. Karimi, H., Roy, P., Saba-Sadiya, S., Tang, J. (2018) Multi-source multi-class fake news detection. In: Proceedings of the 27th international conference on computational linguistics, pp 1546–1557.

19. Wang, W.Y. (2017) Liar, liar pants on fire: A new benchmark dataset for fake news detection. In: Proceedings of the 55th annual meeting of the association for computational linguistics (vol 2: Short Papers), pp 422–426.

20. Kaliyar, R.K., Goswami, A. & Narang, P. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimed Tools Appl 80, 11765–11788 (2021). https://doi.org/10.1007/s11042-020-10183-2.

21. Ashwin, N., Text Classification with Transformers, May 23, 2022. (Available: https://medium.com/@ashwinnaidu1991/text-classification-with-transformers-70acaf65c4a4 )

22. Fake and real news dataset. (Available: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset).

23. He, P., Liu, X., Gao, J., Chen W., "DeBERTa: Decoding-enhanced BERT with Disentangled Attention", 6 Oct, 2021.

24. Vassilieva, N., In Machine Learning, Software Blog, June 22, 2022, "Cerebras Makes It Easy to Harness the Predictive Power of GPT-J". (Available: https://www.cerebras.net/blog/cerebras-makes-it-easy-to-harness-the-predictive-power-of-gpt-j)

25. Ankur, A. P., Apr 30, OpenAI's GPT-3 vs. Open Source Alternatives (GPT-Neo and GPT-J). (Available: https://www.ankursnewsletter.com/p/openais-gpt-3-vs-open-source-alternatives)

26. Radford, A., Wu, J., Amodei, D., Clark, J., Brundage, M., Sutskever, I., "Better Language Models and Their Implications", February 14, 2019.(Available: https://openai.com/blog/better-language-models/)

27. Zeng, Z., Xiong, Y., Ravi, S. N., Acharya, S., Fung, G., Singh, V., "You Only Sample (Almost) Once: Linear Cost Self-Attention Via Bernoulli Sampling", Nov 18, 2021.

28. YOSO, Transformers, Hugging face. (Available: https://huggingface.co/docs/transformers/model_doc/yoso)