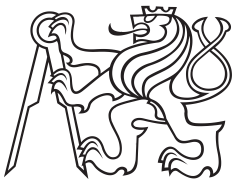
Bachelor thesis



Czech Technical University in Prague

NLP Trolls

Luka Peraica

Supervisor: Ing. Radek Mařík, CSc.
April 2024

Acknowledgements

We thank the CTU in Prague for being a very good *alma mater*.

Declaration

I declare that this work is all my own work and I have cited all sources I have used in the bibliography.

Prague, April 16, 2024

Prohlašuji, že jsem předloženou práci vypracoval samostatně, a že jsem uvedl veškerou použitou literaturu.

V Praze, 16. dubna 2024

Abstract

Keywords: manual, degree project, LaTeX

Supervisor: Ing. Radek Mařík, CSc.

Abstrakt

V záplavě mnoha zdrojů a množství mediálních zpráv není jednoduché se zorientovat i pro profesionální mediální analytiky. Výrazem demokracie je i možnost se ke zprávám vyjadřovat a tříbit si názory v diskusních příspěvcích dílčích zpráv. Diskuse však vytváří prostor i pro osoby, jejichž cílem je z rozmanitých důvodu diskuse narušovat a překrucovat. Cílem práce je vytvořit komponenty systému, který umožní sledovat linie vývoje tématu a identifikovat příspěvky narušitelů, tzv. trollů.

Klíčová slova: manuál, závěrečnná práce, LaTeX

Contents


Figures
Tables

Chapter 1

Introduction

1.1 Problem Statement

In today's flood of diverse media sources and information, even professional media analysts find it challenging to navigate and filter reliable content. A key aspect of democracy is the ability to express opinions and refine perspectives through discussions on news articles. However, these online discussions also create opportunities for individuals whose goal is to disrupt and manipulate conversations for various reasons. The rise of online trolling has become a significant issue, as trolls deliberately provoke, mislead, and incite conflict, thereby spreading misinformation and fostering hostility in digital spaces.

The internet, as a central platform for communication, information sharing, and community building, is increasingly affected by this phenomenon. Studies, such as that by Fornacciari et al.[9], demonstrate that different types of trolls display unique behavioral patterns, emphasizing the need for diverse and adaptive detection methods. Natural Language Processing (NLP) has emerged as a crucial tool in addressing this challenge, offering methods to automatically identify and mitigate the impact of trolls. This thesis aims to develop components of a system capable of tracking the evolution of discussion topics and identifying disruptive contributions from trolls. It provides an overview of various NLP techniques for troll detection, including stylometry, topic modeling, deep learning, and transformer models.

1.2 Structure of the Thesis

Chapter 2

Theoretical Background

2.1 Stylometry

Stylometry is the discipline of analyzing writing style to uncover patterns, identify authors, and extract meaningful details from texts.[10][11] The term was introduced in 1890 by the Polish philosopher Wincenty Lutosławski, who applied it to analyze Plato's works.[7] In the context of this thesis, stylometry involves the use of automated techniques to analyze linguistic traits that distinguish authors based on their unique writing patterns.

One of the core assumptions in computational stylometry is that an author's choices are influenced by sociological factors, such as age, gender, and education level, as well as psychological factors, like personality and native language proficiency.[3] These choices form a distinct, recognizable style that can be analyzed for various purposes, including troll detection. Stylistic features, which play a fundamental role in this process, range from simple surface-level metrics like word length to more complex syntactic and semantic traits.

We can group these features into key categories studied in literature:

- **Lexical Features:** These can be word choices, vocabulary richness, or usage of certain phrases.
- **Syntactic Features:** This involves sentence structure, punctuation usage, and grammatical complexity.[14]
- **Semantic Features:** Which explores meaning and sentiment expressed in a text.[12]

By extracting these features, machine learning classifiers can be trained to recognize troll behavior.

2.1.1 Related Work

An example of stylometry applied to fake news detection is presented in the work of Pérez-Rosas et al. (2018).[12] They used a variety of stylometric features, including n-grams, punctuation frequency, readability metrics and syntactic features. They also incorporated psycholingustic features extracted

from the LIWC lexicon which categorize words into various psychological categories. LIWC features capture psychological aspets of a text such as emotional tone or cognitive processes, potentially revealing underlying psychological differences between fake and legitimate news writers. A linear SVM classifier was trained on these features to differentiate between fake and legitimate news articles. Their results showed that stylometric features can be effective for the task, achieving accuracies of up to 76% which outperformed two human annotators. Their analysis revealed that fake news articles tend to exhibit distinct linguistic patterns compared to legitimate news, such as a greater use of negative emotion words, a greater use of second-person pronouns like (he/she) and a focus on the present.

Though stylometry has proven useful for text classification, recent advancements in large language models and their potential for misuse might pose a challenge to its efficacy in troll detection. As demonstrated by Schuster et al.[15], stylmoetry may struggle to differentiate between human-written and machine-generated text. In their study they find that while a state-of-the-art stylometry-based classifier could effectively detect the presence of machine-generated text within human-written content, it struggled to discern the truthfulness of the generated text. For instance, even a single auto-generated sentence within a longer human-written text was easily detectable, but the veracity of that sentence remained largely undecidable. Additionally, even a relatively weak LM could produce statement inversions that evaded detection by the stylometry-based model.

2.2 Topic Detection Techniques

2.2.1 LDA

Topic detection methods like Latent Dirichlet Allocation (LDA) can analyze online messages, identify common patterns and then group them by topics. LDA is a probabilistic model that identifies latent topics in documents based on word co-occurrence patterns.[2] It assumes that each document is a mixture of topics, and each topic is a distribution over words. By analyzing the distribution of words in troll messages, LDA can uncover the underlying topics that trolls frequently discuss. However, challenges exist in selecting the appropriate algorithm and determining the optimal number of topics.[13] Research has shown that LDA can be effectively used to analyze troll tweets during events like the 2016 US election, revealing coordinated campaigns focused on specific political issues.[4]

2.2.2 Top2Vec

Top2Vec is a state-of-the-art and well-established alternative to traditional topic modeling techniques like LDA. Unlike LDA, where the number of topics needs to be set manually, Top2Vec automatically determines the optimal number of topics. It achieves this by analyzing the density of

document clusters in a vector space created using word embeddings. This is advantageous especially with datasets where the exact number of topics is not clear. Additionally, Top2Vec provides a unique topic representation by identifying the most representative documents and phrases for each topic. This representation, allows for a more nuanced understanding of the topics compared to traditional methods that primarily focus on individual words.[1]

2.2.3 BERTopic

BERTopic is another state-of-the-art technique that can be used for dynamic topic modeling. It leverages pre-trained transformers and Class-based TF-IDF to create dense clusters allowing for easily interpretable topics while keeping important words in the topic descriptions. BERTopic also enables for the analysis of topic evolution by calculating the topic representation at different time step without the need to run the entire model several times.[5]

2.3 Sentiment Analysis

Sentiment analysis is a subfield of natural language processing that focuses on identifying and quantifying the emotional tone behind textual data. By analyzing words and their context it aims to classify text according to its polarity - positive, negative or neutral.

2.3.1 Word2Vec

Word2Vec is a powerful word embedding technique used in NLP to represent words as dense numerical vectors. These vectors capture semantic relationships between words, allowing for meaningful comparisons and analysis. Word2Vec utilizes a shallow neural network to learn word embeddings from a large corpus of text, where the weights of the trained model serve as the embedding vectors. This technique has gained popularity in various NLP tasks, including sentiment analysis. [8]

2.3.2 Related Work

Jiang et al.(2021)[6] explored the use of sentiment analysis for troll detection on the chinese sociel media platform Weibo. They employed a Word2Vec model trained on a dataset of Weibo comments to generate word embeddings. These embeddings were then used to calculate sentiment scores, incorporating features such as happiness, anger, disgust, and fear. The sentiment was used along with meta features such as the location of a comment in a thread or its like count to train XGBoost and SVM models for the troll detection task. The approach proved effective with the XGBoost model achieving an accuracy of up to 89% and SVM up to 87%.

Chapter 3

Appendix A

Bibliography

[1] Dimo Angelov and Diana Inkpen. Topic modeling: Contextual token embeddings are all you need. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13528–13539, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[2] David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. In *The Journal of Machine Learning Research*, volume 3, pages 601–608, 01 2001.

[3] Walter Daelemans. Explanation in computational stylometry. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 451–462, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[4] Hudson Golino, Alexander P. Christensen, Robert Moulder, Seohyun Kim, and Steven M. Boker. Modeling latent topics in social media using dynamic exploratory graph analysis: The case of the right-wing and left-wing trolls in the 2016 us elections. *Psychometrika*, 87(1):156–187, 2022.

[5] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, March 2022. arXiv:2203.05794 [cs].

[6] Zidong Jiang, Fabio Di Troia, and Mark Stamp. Sentiment Analysis for Troll Detection on Weibo. *CoRR*, page 0, March 2021. arXiv:2103.09054 [cs].

[7] Wincenty Lutoslawski. Principes de stylométrie appliqués à la chronologie des œuvres de Platon. *Revue des Études Grecques*, 11(41):61–81, 1898.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[9] Paolo Fornacciari Monica, Mordonini, Agostino Poggi, Laura Sani, and Michele Tomaiuolo. A holistic system for troll detection on Twitter. *Computers in Human Behavior*, 89:258–268, December 2018.

[10] Frederick Mosteller and David L. (David Lee) Wallace. *Inference and disputed authorship: The Federalist.* Reading, Mass., Addison-Wesley, 1964.

[11] Antonio Pascucci, Raffaele Manna, Vincenzo Masucci, and Johanna Monti. The role of computational stylometry in identifying (misogynistic) aggression in english social media text. In *Second Workshop on Trolling, Aggression and Cyberbullying*, 2020.

[12] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[13] Matthias Rüdiger, David Antons, Amol M. Joshi, and Torsten-Oliver Salge. Topic modeling revisited: new evidence on algorithm performance and quality metrics. *PloS one*, 17:e0266325, 2022.

[14] Yunita Sari, Mark Stevenson, and Andreas Vlachos. Topic or Style? Exploring the Most Useful Features for Authorship Attribution. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[15] Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. The limitations of stylometry for detecting machine-generated fake news. 2020.