

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 analytiku. 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 důvodů 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, závěrečná práce, \LaTeX

Contents

1 Introduction	1
1.1 Problem Statement	1
1.2 Structure of the Thesis	1
2 Theoretical Background	3
2.1 Stylometry	3
2.1.1 Related Work	3
2.2 Sentiment Analysis	4
2.2.1 Word2Vec	4
2.2.2 Related Work	4
2.3 Topic Detection	5
2.3.1 LDA	5
2.3.2 Top2Vec	5
2.3.3 BERTopic	6
3	7
A Bibliography	9

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 Fornaciari et al.[10], 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.[11][12] 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.[15]
- **Semantic Features:** Which explores meaning and sentiment expressed in a text.[13]

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).[13] They used a variety of stylometric features, including n-grams, punctuation frequency, readability metrics and syntactic features. They also incorporated psycholinguistic features extracted

from the LIWC lexicon which categorize words into various psychological categories. LIWC features capture psychological aspects 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. The analysis uncovered distinct linguistic patterns in fake news, such as increased use of social and positive words, a focus on present and future actions, and a higher prevalence of adverbs, verbs, and punctuation marks.

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.[16], stylometry 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 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.

2.2.1 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. [9]

2.2.2 Related Work

Jiang et al.(2021)[6] explored the use of sentiment analysis for troll detection on the chinese social 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%. In another paper leveraging sentiment analysis Machova et al. (2022)[8] explored the detection of suspicious reviewers in online discussions, specifically focusing on trolls. Their lexicon-based approach analyzed the polarity of comments to identify trolls. It was based on the tendency of trolls to express extreme opinions that oppose the general sentiment of the discussion. They compared this approach with a Convolutional Neural Network (CNN) model, finding that both performed similarly on text data, achieving accuracies of 0.95 and 0.959, respectively. The study also employed machine learning methods, such as Support Vector Machines (SVM), using non-textual features like comment karma, likes, and dislikes. With the SVM model they achieved an accuracy of 0.986.

■ 2.3 Topic Detection

■ 2.3.1 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.[14] 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]

■ 2.3.2 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]

■ 2.3.3 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]



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] Kristina Machova, Marian Mach, and Matej Vasilko. Comparison of machine learning and sentiment analysis in detection of suspicious online reviewers on different type of data. *Sensors*, 22(1), 2022.
- [9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

