Bachelor thesis

NLP Trolls

Luka Peraica

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.[4], 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.[8][1] 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.[10]

- **Semantic Features:** Which explores meaning and sentiment express in a text.[11]

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

2.2 Topic Detection Techniques

Topic detection methods, such as Latent Dirichlet Allocation (LDA), can be employed to analyze the thematic content of online messages and identify patterns associated with troll behavior. LDA is a probabilistic model

that identifies latent topics in a collection of 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.[9] 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.[5]

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 allows for the analysis of topic evolution by calculating the topic representation at different timestep without the need to run the entire model several times.[6]

2.3 Sentiment Analysis

4

Chapter 3

Appendix A

Bibliography

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

[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] 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.

[5] 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.

[6] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, March 2022. arXiv:2203.05794 [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] Frederick Mosteller and David L. (David Lee) Wallace. *Inference and disputed authorship: The Federalist.* Reading, Mass., Addison-Wesley, 1964.

[9] 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.

[10] 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.

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