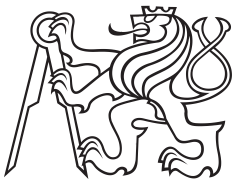
Bachelor thesis



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.[FMP+18], 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. It is commonly used to identify authors and uncover details about their writing style.[MW64] [AP20] The term was introduced in 1890 by Polish philosopher Wincenty Lutosławski, who applied it to analyze Plato's works. [Lut98] In the context of this thesis, stylometry helps quantify distinctive writing patterns that internet trolls exhibit, by extracting features from their text such as:

- 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.[SSV18]
- Semantic Features: Which explores meaning and sentiment express in a text.[ZJ21]

2.2 Sentiment Analysis

2.3 Topic Detection Techniques

2.4

2.5 Troll Detection

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Appendix A

Bibliography

[AP20] Vincenzo Masucci Johanna Monti Antonio Pascucci, Raffaele Manna, *The role of computational stylometry in identifying (misogynistic) aggression in english social media text*, Second Workshop on Trolling, Aggression and Cyberbullying, 2020.

[FMP+18] Paolo Fornacciari, Monica Mordonini, Agostino Poggi, Laura Sani, and Michele Tomaiuolo, *A holistic system for troll detection on Twitter*, Computers in Human Behavior **89** (2018), 258–268.

[Lut98] Wincenty Lutoslawski, *Principes de stylométrie appliqués à la chronologie des œuvres de Platon*, Revue des Études Grecques **11** (1898), no. 41, 61–81 (fre).

[MW64] Frederick Mosteller and David L. (David Lee) Wallace, *Inference and disputed authorship: The Federalist*, Reading, Mass., Addison-Wesley, 1964 (eng).

[SSV18] Yunita Sari, Mark Stevenson, and Andreas Vlachos, *Topic or Style? Exploring the Most Useful Features for Authorship Attribution*, Proceedings of the 27th International Conference on Computational Linguistics (Santa Fe, New Mexico, USA) (Emily M. Bender, Leon Derczynski, and Pierre Isabelle, eds.), Association for Computational Linguistics, August 2018, pp. 343–353.

[ZJ21] Mark Stamp Zidong Jiang, Fabio Di Troia, *Sentiment Analysis for Troll Detection on Weibo*, CoRR (2021), arXiv:2103.09054 [cs].