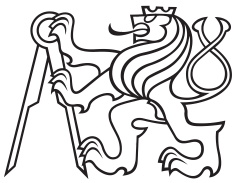
Bachelor thesis



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

The way humans communicate and interact has changed dramatically in the age of the internet. Social media sites, forums and comment sections have become primary spaces for people to share ideas, debate issues and engage in public discourse. These online discussion platforms allow individual from different backgrounds to express their opinions and be part of conversations that shape public perspective more easily than ever before. However, while these platforms create opportunities for connecting people and information sharing, they also come with diverse challenges like the spread of misinformation, polarization and disruptive behavior.

Given these challenges, particularly misinformation and disruptive behavior, it becomes crucial to understand how online discourse shapes public opinion. 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. Social media platforms like Twitter, Facebook and Reddit have a powerful influence on public opinion and can significantly shape political outcomes [BS12]. 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.

1.2 Defining Online Trolling

To address the negative consequences of disruptive online behavior mentioned earlier, it is important to define one its most prevalent forms: online trolling. Online trolling is a deliberate act intended to provoke, deceive, or disrupt online conversations. According to Coles and West [CW16], trolling involves actions meant to annoy, frustrate, or engage others in pointless disputes. Similarly, Golf-Papez and Veer [GPV17] define trolling as "deliberate, deceptive, and mischievous attempts to provoke reactions from other users". While

some forms of trolling may seem harmless or playful, others can escalate into targeted harassment, misinformation campaigns, and efforts to manipulate public opinion.

While understanding trolling as a deliberate provocation clarifies its intent, exploring its underlying psychological motivations can give us further insights into such behavior. People engage in trolling for various reasons, from seeking amusement to pushing political or ideological agendas. Research has shown that certain psychological factors contribute to trolling, such as the "online disinhibition effect". This theory suggests that people act more aggressively online because they feel anonymous and free from real-world consequences [Sul04]. Additionally, studies indicate that personality traits like psychopathy, narcissism, and Machiavellianism are often linked to trolling behavior [BTP14].

Beyond individual psychology trolls also exploit broader technological factors, particularly social media algorithms that prioritize engagement. Effectively playing into the algorithm allows them to more easily and effectively spread divisive content and manipulate conversations [GPV17].

1.3 Societal and Political Impacts of Trolling

Trolling negatively affects honest individuals involved in online discussions. Those targeted by trolls often experience stress, anxiety, and frustration, which can discourage them from participating in online discourse. Repeated exposure to trolling can drive individuals away from digital platforms, silencing voices that would otherwise contribute to meaningful discussions. This type of behavior not only harms individual well-being but also degrades the quality of discussions. As user trust is eroded and people become more skeptical of digital interactions a toxic environment is created where constructive engagement becomes difficult [GPV17].

On a larger scale trolling has significant consequences, particularly when it is used as a toll for political manipulation. State-sponsored troll campaigns have been used to spread propaganda, influence elections, and undermine public trust in media [BH17]. One of the most well-known exaples is the Russian *Internet Research Agency* (IRA), which ran large-scale trolling operations during the 2016 U.S. presidential election between Hillary Clinton and Donald Trump. These trolls used fake accounts to post divisive content and manipulate public discourse [LW20]. Similar use of trolling in political campaigns and foreign influence operations has been documented across the world, demonstrating the severity and importance of addressing the issue.

This thesis aims to contribute to the fight against trolling by developing methods to track how discussions evolve and identifying harmful contributions. Specifically, it will explore different NLP techniques for troll detection, including stylometry, topic modeling, deep learning, and transformer models.

Chapter 2

Related Methods

Given the impact of disruptive trolls on online discourse and society at large, research efforts have focused on developing techniques to better understand, detect and mitigate their activity. This chapter explores the methods used to analyze and identify trolling behavior particularly through Natural Language Processing (NLP). It covers key approaches such as stylometry, sentiment analysis, and topic modelling.

2.1 Stylometry

Stylometry is the discipline of analyzing writing style to uncover patterns, identify authors, and extract meaningful details from texts [MW64] [PMMM20]. The term was introduced in 1890 by the Polish philosopher Wincenty Lutosławski, who applied it to analyze Plato's works [Lut98]. 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 [Dae13]. 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 [SSV18].
- **Semantic Features:** Which explores meaning and sentiment expressed in a text [PRKLM18].

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. [PRKLM18]. 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. 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.

In another paper, Kandasamy et at. [KTM+21] proposed a deep learning framework for sentiment analysis of COVID-19-related tweets. Their approach used an N-gram stacked autoencoder to capture text features. These features were then processed by a set of classifiers—decision trees, support vector machines, random forests, and k-nearest neighbors. The highest accuracy was achieved using an ensemble model that combined all of these classifiers, this method achieved an accuracy of 87,75%. The study demonstrated that using n-grams greatly improved the classification of negative sentiment, an emotion that was prevalent during the pandemic.

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. [SSSB20], 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 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 [MCCD13]. 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.

2.2.2 Related Work

Jiang et al. [JTS21] 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. [MMV22] 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) [BNJ01] 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. 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 [RAJS22]. 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 [GCM+22].

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 [AI24].

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 [Gro22].

Chapter 3

Experiments

3.1 Dataset

The dataset used in this thesis consists of user-generated comments collected from the discussion sections under news articles published on Novinky.cz, one of the largest Czech news portals. Each article on Novinky.cz includes a public comment section where users actively engage in discussions about the content presented. These discussions are often extensive, with some articles attracting hundreds of user comments.

In the Czech online media landscape, it is widely recognized that the comment sections on major news sites, particularly on Novinky.cz for example, frequently serve as hotbeds for controversy and emotionally charged discourse. They are often perceived by the public as spaces where individuals express grievances, frustrations, and polarizing viewpoints, sometimes in ways that border on or cross into what could be described as abusive, manipulative or troll like behavior. This cultural context makes Novinky.cz a relevant and interesting setting for exploring how online discussions develop, especially where conversations become heated or emotionally charged.

For the purposes of this thesis, a large-scale dataset comprising approximately 350,000 comments posted by around 48,000 users was provided by Newton Media, a prominent media intelligence organization.

Each data entry includes the following attributes:

- **Comment content** – the full textual body of the comment.
- **Article metadata** – including the title and link to the article under which the comment was posted
- **Timestamp** – the date and time when the comment was published.
- **Author name** – full author name as collected from the discussion.
- **Sentiment label** – a sentiment category assigned by Newton Media, labeled as one of the following: *Neutral*, *Positive*, *Negative*, or *Ambivalent*.

A key challenge posed by this dataset is the absence of explicit troll/non-troll labels. Since there is no ground-truth annotation for trolling behavior, we

cannot directly apply supervised classification methods. Because of this, we have to rely on unsupervised or semi-supervised techniques, such as clustering, topic modeling, or anomaly detection, to try to find patterns that may point to trolling based on how the comments are written and their sentiment. The lack of labeled data also makes it difficult to verify the accuracy or effectiveness of any classifications or patterns identified during experimentation. Without labeled data, we cannot easily measure how accurate our models are, and must instead rely on manual checks, indirect indicators, or interpretation of the patterns found

Appendix A

Bibliography

[AI24] 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.

[BH17] S Bradshaw and P Howard. Troops, trolls and troublemakers: A global inventory of organized social media manipulation. Technical report, 2017.

[BNJ01] David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. In *The Journal of Machine Learning Research*, volume 3, pages 601–608, 01 2001.

[BS12] Lance Bennett and Alexandra Segerberg. The logic of connective action: Digital media and the personalization of contentious politics. *The Logic of Connective Action: Digital Media and the Personalization of Contentious Politics*, 15:1–240, 01 2012.

[BTP14] Erin E. Buckels, Paul D. Trapnell, and Delroy L. Paulhus. Trolls just want to have fun. *Personality and Individual Differences*, 67:97–102, 2014. The Dark Triad of Personality.

[CW16] Bryn Alexander Coles and Melanie West. Trolling the trolls: Online forum users constructions of the nature and properties of trolling. *Computers in Human Behavior*, 60:233–244, 2016.

[Dae13] 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.

[GCM+22] 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.

[GPV17] Maja Golf-Papez and Ekant Veer. Don't feed the trolling: rethinking how online trolling is being defined and combated. *Journal of Marketing Management*, 33(15/16):1336–1354, November 2017.

[Gro22] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, March 2022. arXiv:2203.05794 [cs].

[JTS21] Zidong Jiang, Fabio Di Troia, and Mark Stamp. Sentiment Analysis for Troll Detection on Weibo. *CoRR*, page 0, March 2021. arXiv:2103.09054 [cs].

[KTM+21] Venkatachalam Kandasamy, Pavel Trojovský, Fadi Al Machot, Kyandoghere Kyamakya, Nebojsa Bacanin, Sameh Askar, and Mohamed Abouhawwash. Sentimental analysis of covid-19 related messages in social networks by involving an n-gram stacked autoencoder integrated in an ensemble learning scheme. *Sensors*, 21(22), 2021.

[Lut98] Wincenty Lutoslawski. Principes de stylométrie appliqués à la chronologie des œuvres de Platon. *Revue des Études Grecques*, 11(41):61–81, 1898.

[LW20] D. Linvill and Patrick Warren. Troll factories: Manufacturing specialized disinformation on twitter. *Political Communication*, 37:1–21, 02 2020.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[MMV22] 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.

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

[PMMM20] 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.

[PRKLM18] 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.

[RAJS22] 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.

[SSSB20] Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. The limitations of stylometry for detecting machine-generated fake news. In *Booktitle*, 2020.

[SSV18] 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.

[Sul04] John Suler. The Online Disinhibition Effect. *CyberPsychology & Behavior*, 7(3):321–326, June 2004.