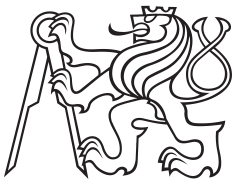
Bachelor thesis



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 easily than ever before. However, while online discussions create opportunities for connecting people and sharing information, they also come with major challenges like the spread of misinformation, the polarization of society and the spread of large scale disruptive behavior.

Understanding how these platforms shape opinion is essential in the modern day. As in today's flood of diverse media sources and information, even professional media analysts find it challenging to navigate and filter reliable information. A key aspect of democracy is the ability to express opinions and refine perspectives through discussions. Today most of such discussion happens online on social media platforms like Twitter, Facebook and Reddit which have gained a powerful influence on public opinion the shaping political outcomes [BS12]. This importance and ubiquity of online discussions can also make them targets for individuals whose goal is to disrupt and manipulate conversations for various reasons.

1.2 Defining Online Trolling

To address the negative consequences of disruptive online behavior, 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".

The term "trolling" itself was originally borrowed from fishing slang, where it referred to dragging a baited line through the water to catch fish. In the

online context, the term seems to have first been used in the 1990s on the USET discussion system where some users would deliberately create posts designed to trigger angry corrections from newbie users who weren't aware of such pracitces.

Then there is a second definition for the word "troll", which is also quite relevant to the perception of online trolls and perhaps for most people the first connotation that comes to mind. This definition refers to a troll as a large, ugly creature from folklore, often depicted brutish ogre. The word "troll" is derived from the Old Norse word "troll", which means "giant" or "ogre". In this context, the term evokes an image of a monstrous being that lurks in the shadows, waiting to pounce on unsuspecting victims. And while the term trolling originated from the early bait posts, related to the fishing term, over time the the character and label of the "troll" developed, which is more closely related to the folklore definition. This shift in meaning reflects the evolution of online trolling from more light-hearted baiting and joking to a label for a malicious character lurking on the internet.[DBGJS21]

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

People engage in trolling for various reasons, from simply seeking amusement from the activity to pushing political or ideological agendas. Studies indicate that personality traits like psychopathy, narcissism, and Machiavellianism are often linked to trolling behavior [BTP14]. Additionally research has shown that certain psychological factors also contribute to the online trolling phenomena, 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]. The combination of these factors makes online discussions particularly fertile ground for antisocial behavior.

Beyond individual psychology trolls also exploit the techonological factor of the discussions, particularly social media algorithms that focus on engagement above all else. Effectively playing into the algorithm allows them to more easily and effectively spread divisive content and manipulate conversations [GPV17].

1.3 Impacts of Trolling

Trolling negatively affects both honest users and the discussion as a whole. Those targeted by trolls often experience stress, anxiety, and frustration, which can discourage them from engaging in future conversations. Trolling 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 can have 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 identify and analyze behavior of trolls in online discussions. Specifically, it will explore different NLP techniques for troll detection, including stylometry, topic modeling, deep learning, and transformer models. The goal is to identify harmful contributions and contributors to online discussions and to explore possibilites for further research in this area.

Chapter 2

Natural Language Processing

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.

The underlying assumption in 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]. This assumption can be extended to groups of authors, especially those who may share common objectives or adhere to specific guidelines, such as state-sponsored trolls, or display similar behavioral patterns as seen among ordinary trolls. These individual or collective choices can manifest as identifiable stylistic features within texts, which computational models can analyze to detect trolling behavior. Stylometric analyses typically examine lexical choices like vocabulary richness, syntactic elements including sentence structure and grammatical complexity [SSV18], and semantic dimensions, such as sentiment and thematic consistency [PRKLM18]. Extracting and evaluating these features allows machine learning classifiers to differentiate between regular users and trolls based on their distinctive linguistic signatures.

2.1.1 Stylometry in Literature

An example of stylometry applied to troll detection is presented in the work of Machová *et al.* [MPH21]. The paper examines troll detection in Slovak Facebook discussions on COVID-19 by combining shallow stylometric cues with engagement and affective information. From roughly 2,500 manually labelled comments they extract length-based metrics (character and word counts, average word length), orthographic signals (capital-letter and digit frequency), and eight sentiment/provocativeness categories, and enrich these with interaction metadata such as the number of "likes." Classical classifiers-including SVM, Multinomial Naïve Bayes (MNB) and logistic regression-are trained on bag-of-words and TF-IDF representations. The MNB model that integrates stylometric, affective and metadata features achieves the best balance, reaching 0.92 recall for the troll class, while an SVM attains perfect precision (1.00) at the cost of markedly lower recall. Their results confirm that stylistic signals are informative but deliver the highest performance when fused with complementary sentiment and platform-level features, rather than being relied upon in isolation.

In another paper 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.

Though stylometry has proven useful for text classification, recent advancements in large language models and their potential for misuse might pose a substantial challenge to its efficacy. 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.

These findings collectively highlight stylometry's potential for detecting hidden manipulation in online text, although recent advancements in language

generation models present new challenges. It is also imporatnt to note that to achieve good results stylometric features weren't used on their own but along with others like meta data (like counts, followers) or sentiemnt analysis.

2.2 Sentiment Analysis

Sentiment analysis is a method of determining the emotional tone of text, which can be achieved using lexicon-based methods, machine learning, or deep learning approaches. It identifies positive, negative, neutral, or ambivalent tones in text.

2.2.1 Sentiment Analysis in Troll Detection

The use of sentiment analysis has been explored as one of the methods used of troll detection. Jiang et al. [JTS21] for example 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 a related study, Machová et al. [MMV22] investigated the detection of suspicious reviewers in online discussions, 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

Topic detection is another essential NLP technique used to analyze and interpret corpora of text. It identifies main topics in large amounts of text and groups conversations based on these topics. This can help us understand conversation patterns and recognize signs of disruptive behavior. Trolls often move discussions off-topic, introduce controversial subjects, or focus repeatedly on divisive issues. We can analyze the kinds of topics a user engages with and how they behave within those topics to find clues about their intentions and their role in discussions.

2.4 Transformer Models

In recent years, Transformer models have emerged as the state-of-the-art approach for a wide range of NLP tasks. Introduced by First introduced by Vaswani et al. in 2017 [VSP+17], Transformers offer a new way for machines to understand and generate language. Unlike earlier methods, they are designed to efficiently capture relationships and patterns within text, even over long passages. Their ability to handle large amounts of data and to adapt to complex language structures has made them a key tool in modern NLP applications. Another important reason for the success of Transformer models is that they are typically very large models pre-trained on massive datasets. Many of these datasets are multilingual, meaning that the models learn from several languages at once. As a result, we can often use the same embedding space for different languages, and the knowledge gained in one language can transfer to some extent to others. This makes pre-trained Transformers especially valuable when dealing with multilingual or cross-lingual data.

In this section, we will introduce the basic ideas behind Transformer models.

2.4.1 Self-Attention

The key innovation of Transformer models is the use of self-attention mechanisms, which allows the model to dynamically focus on different parts of the text sequence, regardless of position. Instead of processing text word by word like earlier models, Transformers can consider all words in a sentence at once, deciding how much attention each word should pay to others. This allows the model to capture complex patterns and dependencies across long texts, making it very powerful for understanding the both the context and meaning of language.

Each word (token) creates three vectors which are used by the self-attention mechanism:

Query (Q) - a vector that represents the active word in the input sequence ("What am I looking for?").

Key (K) - a vector that represents other words form the input sequence which are compared to the query ("What do I have?").

Value (V) - a vector that represents the information we want to extract from the compared words ("What do I want to know?")

The model uses a dot product to calculate the similairy between the query and the key vectors. The result is then scaled and passed through a softmax function to create a probability distribution. This distribution is then used to weight the value vectors, allowing the model to focus on the most relevant information in the input sequence. The model does all of the above steps in paralel for sets of queries, keys and values - it works with matrices instead of separate vectors. The equation for the self-attention mechanism as described is then:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

2.4.2 Multi-Head Attention

Additionally to the self-attention mechanism, Transformer models also use multi-head attention. This means that instead of having a single set of query, key and value vectors, the model has multiple sets (or heads) that can learn different aspects of the input data. Each head performs its own self-attention calculation, and the results are then concatenated and linearly transformed to create the final output. This allows the model to capture a wider range of relationships and patterns in the data.

The multi-head attention mechanism is defined as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O \tag{2.2}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.3}$$

The output is finally passed through a feed-forward neural network (FFN) which consists of two linear transformations with a ReLU activation function in between. The FFN is applied to each position separately and identically.

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{2.4}$$

2.4.3 BERT

One of the most influential Transformer-based models for natural language understanding is BERT (Bidirectional Encoder Representations from Transformers). Introduced by Devlin et al. in 2018 [DCLT18], BERT builds on the Transformer encoder architecture and is pre-trained on large text corpora using masked language modeling and next sentence prediction tasks. Unlike traditional language models that read text from left to right, BERT is deeply bidirectional, meaning it learns information from both the left and right context at the same time. This bidirectional training allows BERT to capture richer information about language, making it highly effective for downstream tasks like text classification, question answering, and sentiment analysis.

2.4.4 Transformer Models in Troll Detection

The paper MetaTroll proposed MetaTroll, a few-shot troll detection framework designed to adapt quickly to new state-sponsored influence campaigns using minimal labeled data. Their approach is based on a meta-learning framework and incorporates campaign-specific transformer adapters to tackle catastrophic forgetting, a common problem where models lose the ability to detect trolls from older campaigns after continual updates. MetaTroll outperformed traditional n-gram SVM baselines, achieving a 92.3% F1-score in

5-shot scenarios and demonstrating strong cross-lingual capabilities.[TZL23] Embeddings generated by transformer models have been shown to outperform static methods due to their ability to capture the contextual meanings of words. BERT based word embeddings outperformed static GloVe vectors are reached AUC scores of 0.924. The authors of the paper [YZ23] highlight how transformers can better cepture contextual nuances in language, which can be important for the task of troll classification, where language can be manipulative and deliberately deceptive.

In this thesis, we will attempt to leverage pre-trained Transformer such models as BERT and try to fine tune them for the troll detection task. The reasons for the choice of using Transformers will be outlined in the next chapter.

Chapter 3

Dataset

Before outlining the proposed method, I will first describe the dataset that forms the basis of our work. The properties of the dataset are important as it directly shape the choice of methods and defines the limitiations and the goals of the work.

3.1 Main 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 participate in discussions about the content presented. These discussions are often extensive, with some articles attracting hundreds of user comments.

In the Czech online media environment, it is generally 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 consideration for the work with this dataset is the nature of the comments. The comments in general seem to be mostly negative in tone and often emotionally charged, as can be seen by the distribution of setiment.

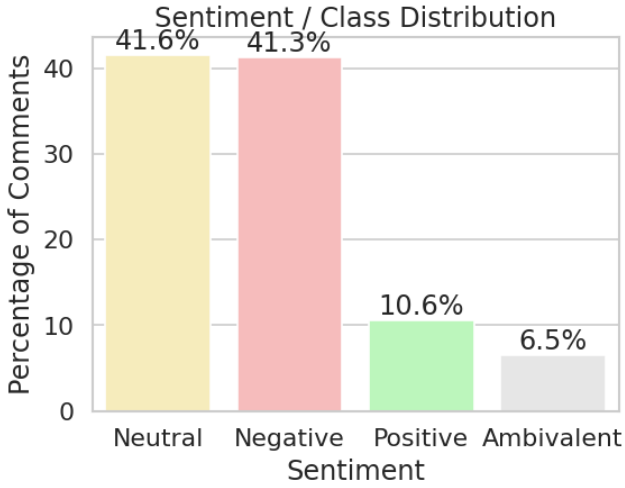


Figure 3.1: Sentiment distribution of collected novinky.cz comments

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 have to instead rely on manual checks and interpretation of the results.

3.2 Additionall Datasets

In addition to the primary dataset collected from Novinky.cz, several publicly available labeled datasets were also used in this thesis. They were collected from different platforms such as Twitter or Reddit and the majority are in English, although some include other languages as well. While they differ in context and language, they offer labeled examples that usd for pre-training of models for the later analysis of Czech online discussions.

3.2.1 IRA Troll Tweets

One of the external datasets used in this thesis is a collection of tweets linked to the Russian *Internet Research Agency* (IRA), which was menitioned

in the introductory chapter of the thesis. The dataset was published by FiveThirtyEight in connection with their article *Why We're Sharing 3 Million Russian Troll Tweets*, and was originally collected by researchers from Clemson University. It contains nearly 3 million tweets posted between February 2012 and May 2018 by accounts identified by Twitter as being linked to the IRA, which were provided to the US Congress for investigation into 2016 presidential election interference. In total, the dataset includes 2,973,371 tweets from 2,848 Twitter handles. Most of the tweets are in English, but some are in other languages, including Russian or German. The dataset is publicly available and can be accessed through the FiveThirtyEight GitHub repository.[1]

3.2.2 Information Operations Dataset

Another external resource used in this thesis is a collection of labeled datasets for research on information operations (IOs), introduced by Seckin et al. [SPN+24]. The full collection contains over 13 million posts from approximately 303,000 accounts. The dataset includes both verified IO posts and control data from legitimate accounts, covering 26 distinct manipulation campaigns originating from different countries. The data is organized first by one of 16 identified state actors, such as Russia, China, or even Catalonia, and then further subdivided into distinct operations. The IO posts were identified and released by major social media platforms including Twitter, Facebook, and Reddit, while the control data captures organic user discussions on similar topics within the same time frames.

3.2.3 Non-troll Datasets

In addition to the troll datasets several non-troll datasets were also collected to ensure availability of apolitical and organic user discussion.

The first non-troll dataset is the Civil Comments dataset, obtained from Hugging Face. It consists of public comments posted between 2015 and 2017 on approximately 50 English-language news sites. Each comment is labeled with values for toxicity, obscenity, and other attributes. For the purposes of this thesis, only comments labeled as non-toxic (toxicity score between 0 and 0.1) were used, which represent the majority of the dataset.[2]

Additionally, a small dataset of celebrity tweets was obtained from Kaggle. This dataset consists of posts by well-known public figures providing examples of casual and generally non-political online communication.[3].

Finally, a custom dataset was manually created by scraping tweets from Czech public figures and politicians. A selected list of Twitter accounts was compiled, and 20 tweets were collected from each account.

[1] `https://github.com/fivethirtyeight/russian-troll-tweets/`
[2] `https://huggingface.co/datasets/google/civil_comments`
[3] `https://www.kaggle.com/datasets/abaghyangor/celebrity-tweets`

Chapter 4

Proposed Method

In this chapter I will outline the proposed method for detecting troll-like behavior in online discussions. The core idea behind the method is the use of transformer-based models, specifically multilingual BERT based models, in a regression task designed to quantify the a users troll-like behavior. Instead of a binary classification task, the approach is to assign a user with a continous "trolliness" score, measured from 0 to 1.

4.1 Motivation

As the backbone of the method, I decided to use multilingual BERT based models, as they are trained across dozens of languages at once, which makes them a natural choice when trying to transfer knowledge from English or Russian troll datasets to Czech. Beyond their multilingual capabilities, BERT models are also able to capture and represent both syntactic and semantic relationships and dependencies within a text sequence. Instead of manually designing and extracting individual features like syntax counts, stylometric traits, sentiment scores, in theory BERT should be able to learn and encode much of this information into its embeddings and attention mechanism.[RKR20]

A classical machine learning approach using stylometric and other features is not suitable for this task, due to the limitations of the datasets we are working with, which were mentioned above. However BERT should be able to capture similar semantic and syntactic knowledge while also being able to be used in our specific task with limited labeled data and multilingual datasets.

The motivation to use a regression task instead of a binary classification task is twofold. First, the main dataset of Czech comments lacks troll/non-troll labels, so standart supervised classification methods cannot be applied. Second, troll behavior isn't a straight forward binary state, but rather a spectrum of behavior, with users displaying varying degrees and different types of distruptive behavior.

4.2 Data Collection and Preprocessing

The first step of the method is the collection and preprocessing of the data. The raw text data is cleaned and preprocessed using basic text preprocessing techniques to normalize it to a certain extend across the different data sources. Each comment is then grouped according to its author, creating sets of comments for each user.

A key design decision in this work was to rate the trolliness at the user level, rather than at an individual comment level. This decision was based on the analysis and observations from the labeled troll datasets. A reccuring pattern was that many troll accounts did not only engage in disruptive and manipulative behavior all the time. Instead, in many cases trolls posted mostly ”normal” content, perhaps to blend in with regular users, pushing their agenda more subdly in some posts and then only occasionally posting more overtly troll-like comments.

For this thesis we will exclude all users with fewer than 5 comments, as our aim is to try to find broader patterns of troll-like behavior not only one-off examples of offensive or provocative comments. We do this both for the initial training as well as when working with the target Czech dataset. While this discards about half of the users in the dataset, it is only a small fraction of the comments, about ten precent.

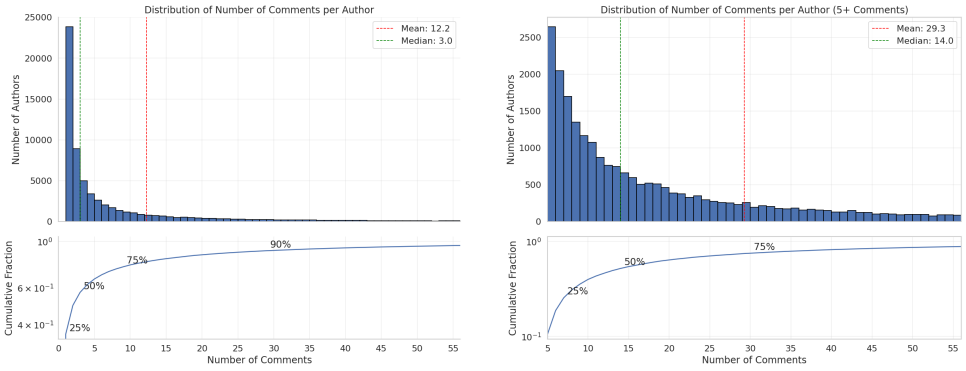


Figure 4.3: Distribution of comments per autor before and after filtering for 5+ comments

4.3 Model Architecture

The model architecture is designed in two levels: the comment level and the user level. At the comment level, each individual comment is first encoded using a pretrained multilingual BERT model. BERT processes the comments and produces fixed-length embeddings which capture both syntactic and semantic information.

At the user level, the embeddings of all the comments from a single user are aggregated. To combine all the comments into a single user-level vector, an

attention mechanism is used. The attention mechanism allows the model to assign different importance to different comments, depending on how strongly they contribute to the the users trolliness. This allows the model to give greater influence to models which are more disruptive or otherwise suspicious when creating the final user representation.

Finally, a regression head is applied on top of the aggregated user-level embedding vector. The regression head consists of a feed-forward neural network, which outputs a the final coutninous trolliness score between 0 and 1. This score should reflect how similar a users behavior is to that of known trolls, rather than to give a hard classification.

4.4 Training

The training of the model is done in two steps. Larger training on the large labeled troll datasets from foreign domains, and a smaller fine-tune on manually annotated Czech comments from the target dataset.

The first training step includes the Russian IRA troll tweets, information operations datasets, and the non-troll datasets like Civil Comments. The training is done using a regression objective, where the model is trained to predict the trolliness of the users instead of their binary class.

Since the labeled training data comes from different domains and languages than our target Czech dataset, a second small fine-tuning step is performed.

After the initial training, the model is fine-tuned on a small set of manually annoated Czech user comments from our target dataset. This data was created by me, by exploring the users in the who were classified with high or low trolliness scores and high confidence during preliminary runs. This few-shot tuning step helps the model better adapt to our specific domain.

4.5 Scoring

Once the model is trained, it can be used to score the trolliness of users in the target dataset. For each user, all available comments are collected and grouped into batches of a fixed size. If a user has fewer comments than needed to fill a batch the comments are padded with empty comments. Each batch of comments is then passed through the model to generate a score. When a user has multple batches, the final trolliness score is calculated as the average over all batches.

The output of the model is then a continous trolliness score between 0 and 1. A binary predition can be obtained by applying a threshold. Additionally the model provides attention weights for individual comments, which how much each comment contributed to the final score and can be used for analysis and interpretation.

4.6 Lightweight Annotation and Evaluation App

Additionally, to help me work with the model and dataset, I created a simple annotation application. The main goal of the tool is to allows labeling of users from the Czech dataset and search for their comments. The app also shows the model's predicted trolliness scores and attention weights. The annotations can then be used both for few-shot fine-tuning and for manual exploration of the model's predictions.

The app loads a saved model checkpoint and available Czech comments from the dataset. It then allows a user to search for an author by name. It then displays the predicted trolliness score of the author and all of their comments along with their attention weights. Finally, the app allows the user to label the author as troll, non-troll or uncertain and saves the labels to a file. The app is implemented in Python using the Streamlit library, which allows for easy creation of interactive web applications.

We used the app to manually label a small set of users from the Czech dataset. We focused on users with high or low trolliness scores, as well as those with uncertain scores. The task proved to be quite challanging. Most users tended to be quite negative and angry in general as mentioned in the previous section, but it was still difficult to rate them as troll.

The manual labelling highlighted the inherent difficulty of the task of distinguishing between toxic behaviour and geinuine disagreement in online discussion. The challange of recognizing trolls from other disruptive or even geniune but negative users could be challangeing even for experts in political science or psychology and goes beyond the scope of this thesis. Depsite the complexity and the time-consuming nature of the annotation we created a small labeled dataset for further use.

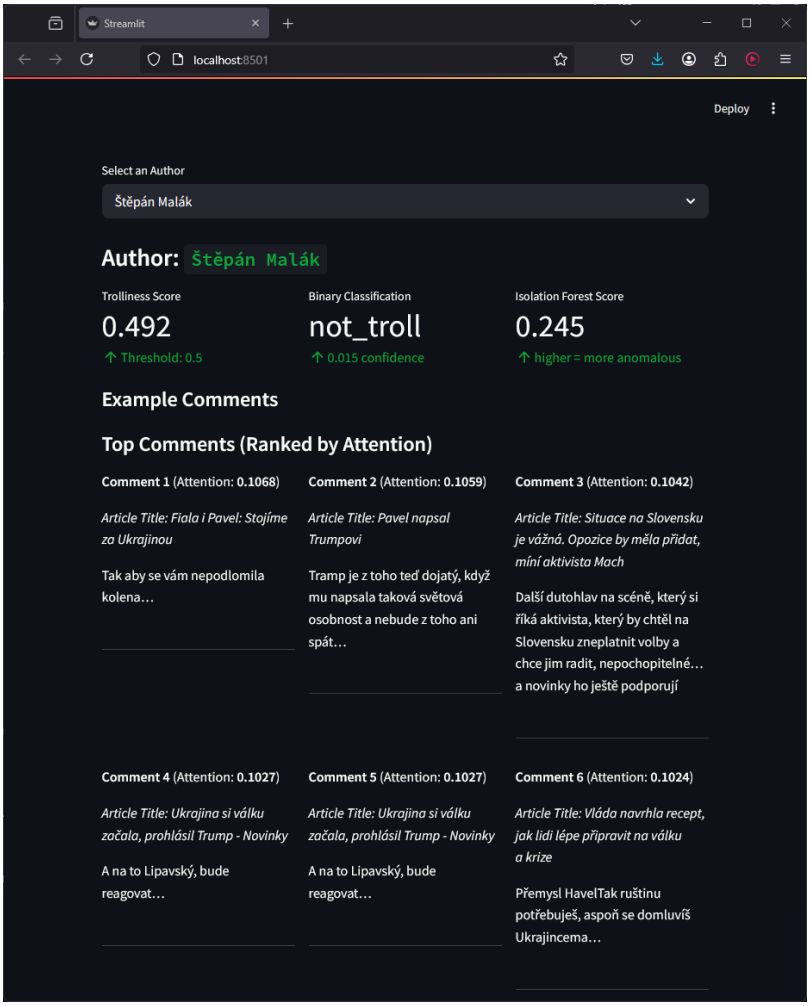


Figure 4.4: Annotation app interface

Chapter 5

Experiments

5.1 Baseline Comparison

To evaluate the reasonability of the proposed method and how much a transformer model really adds, we first reproduced a stylometry-only baseline with the same train/test/validation data splits used with the BERT model. We used 5 stylometric features named in [MPH21] - character count, word count, average word length, capital-letter ratio and digit ratio. We then trained two regressors on them: a linear support-vector regressor (SVR) and a gradient-boosting regressor (GBR). We also added a stronger Term Frequency - Inverse Document Frequency (TF-IDF) based baseline, where comments were represented through 50000 unigrams and bigrams, and fitted on a ridge regressor.

For the BERT model, we used a regression head with a Sigmoid activation function, allowing it to output a continuous score between 0 and 1, which we interpret as a trolliness score. Although the model is structured as a regressor, we used Binary Cross-Entropy Loss (BCE Loss) during training. This choice ensures that the model learns to produce outputs that behave like probabilities - values closer to 1 indicate high trolliness, while values closer to 0 indicate low trolliness. BCE Loss is defined as follows:

$$L_{BCE}(y, \hat{y}) = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right] \tag{5.1}$$

The decision to use BCE Loss, despite framing the problem as a regression task, was motivated by its ability to directly optimize for the probability of each label (0 or 1). Unlike standard regression losses (like Mean Squared Error or Huber Loss), BCE Loss is specifically designed to handle cases where the target values are binary, but the model's predictions are continuous probabilities. This setup helps the model avoid becoming overly biased towards low values, which was a problem with a test run with HuberLoss.

We will compare the training results of the models through their mean square error (MSE) and the coefficient of determination (the R^2 score). The mean square error (MSE) is a measure of the average squared difference between the predicted and actual values. A lower MSE indicates a better fit

of the model to the data.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.2}$$

The R^2 score is a statistical measure of what share of the original variance of the data the model explains. To explain how it works let's first define the equations. First we define two sums of squares:

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5.3}$$

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.4}$$

The R^2 score is then:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{5.5}$$

The R^2 score quantifes the proportion of total variance (TSS) that is left unexplained by the model's residuals (RSS). A value of 1 indicates a perfect fit and 0 indicates that the model does not explain any of the variance in the data. A negative R^2 score indicates that the model performs worse than simply guessing the mean of the target variable.

5.1.1 Evaluation of the Baseline

With coarse stylometric features the SVR and GB regressors did not beat guessing the mean with a negative R^2 score and a MSE of about 0.26. This result is not too surprising as 5 features is much for the regression models to work with, but it shows notheless that few handpicked stylometric features are not enough. Extending to 50000 features (uni- and bi-gram counts) through TF-IDF improves R^2 score a lot, to 0.57 and has a mean square error of 0.09 but it still gets outperformed by the BERT model with a regression head which achieved a R^2 score of 0.65 and a MSE of 0.07. The results of the two best models TF-IDF and BERT are similiar, but the trained TF-IDF model cannot be used on other other language domains, whie the BERT model should be able to carry over some of its learned knowledge. The results of the model training and evaluation are shown in the table below.

Model	Features	Val MSE	Val R^2	Test MSE	Test R^2
Linear SVR	5 stylometric	0.260	−0.211	0.260	−0.211
Ridge TF–IDF	50 000 1-2-grams	0.090	0.579	0.097	0.547
DistilBERT	contextual	**0.087**	**0.592**	**0.082**	**0.618**

Table 5.1: Author-level results on English train/test/val data.

We thus show that BERT achieves similar or better results than a stylometric model and we validate why we chose to use it.

5.2 Further Experiments with BERT Model

For further experiments we trained the BERT model on a mix of all of the availalbe troll comments not only the english language ones like in the baseline evaluation step. Two limitations were made to reduce the size of the training dataset slightly. First, wile loading the data each author was limited to a maximum of 50 comments. Secondly, after loading all the datasets as the full dataset was heavily biased towards the non troll class with 40000 accounts and only 4000 troll accounts, and a ratio of 90% to 10% troll and non troll comments the dataset was balanced to a 50% 50% comment ratio and about a 66% 33% user ratio. The final training data split was as follows:

Category	Count	Percentage
Troll tweets	144,563	51.1%
Non-troll tweets	138,340	48.9%
Total tweets	282,903	100%

Table 5.2: Tweet distribution in the dataset.

Category	Count	Percentage
Troll authors	4,555	35.6%
Non-troll authors	8,236	64.4%

Table 5.3: Author statistics in the dataset.

Dataset	Samples	Authors
Train	198,385	8,953
Validation	42,099	1,919
Test	42,419	1,919

Table 5.4: Dataset sizes and author distribution.

The model was trained for 5 epochs on a single NVIDIA RTX 3060ti GPU (8GB memory) using a batch size of 16 and the AdamW optimizer with a learning rate of 1e-5 and weight decay of 0.01. A linear learning rate schedule was employed. The training, utilizing PyTorch and the HuggingFace Transformers library for the BERT model, took approximately 4 hours. The table and figure below show the training results, including training/validation loss and R^2 scores over epochs.

5.2.1 Predictions on Czech Dataset

After training fine-tuning the BERT model with our prepared data the next step is to explore the model behaviour on the Czech comments domain. The

Metric	Training	Validation
Loss	0.5437	0.5557
Mean Squared Error (MSE)	0.0734	0.0746
R^2 Score	0.6814	0.6641
Binary Accuracy	0.9134	0.9165

Table 5.5: Training and validation metrics at the best epoch (Epoch 2).

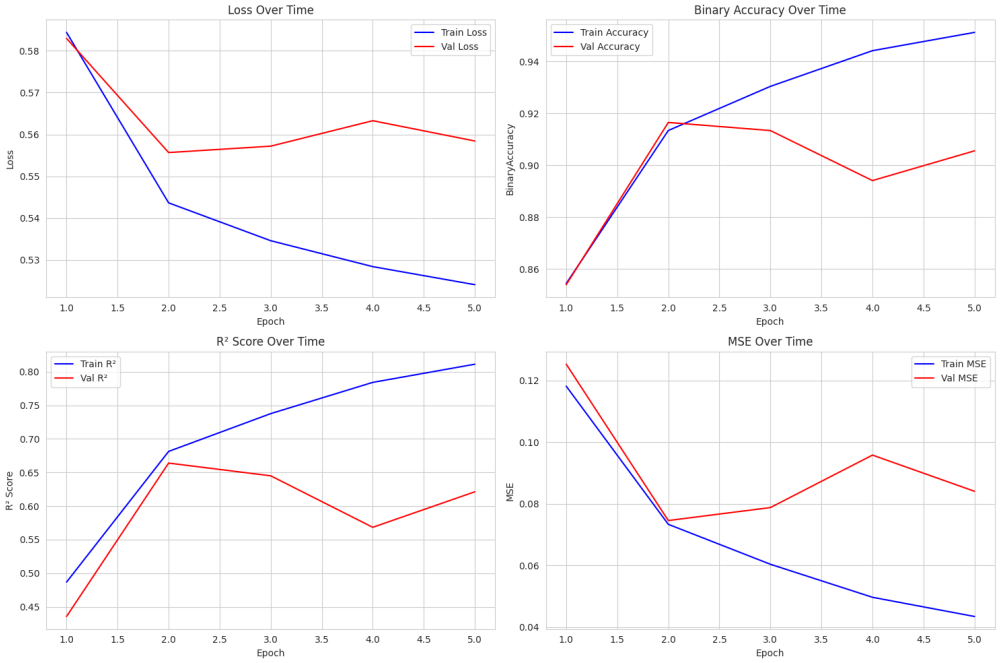


Figure 5.1: Training and validation loss over epochs

goal of the experiemnts is to compare different training strategies and see how they do with trolliness scoring of the Czech dataset.

I compare two different versions of the model.

- **BERT + Troll Training** - DistillBERT fine tuned with foreign troll datasets as described above.

- **BERT + Troll Training + Few Shot Czech Fine-Tune** - model further fine-tuned on a small manually annotated set of Czech users.

Each model is evaluated based on the distribution of predicted trolliness scores across users, manual review of selected user cases and performace on a small hand annotated test set.

5.3 Results

5.3.1 First Experiment

The first set of experiments involved trying to directly plug in the Czech comments into the **BERT + Troll Training** model, which was fine-tuned on the international troll datasets.

The results of this initial approach were concerning. The model exihibited a strong tendency to predict very low trolliness scores across all users, with the majority of scores falling bellow 0.1. The summary statitstics of these predictions are shown in the table below.

Statistic	Value
Mean	0.00019
Standard Deviation	0.00572
Median	0.00003
Maximum	0.41994

Table 5.6: Trolliness score distribution.

These results show that the model struggles to meaningfully capture variation of trolliness in the Czech comments and wasn't able to generalize well to the Czech domain. There could be several factors contributing to this issue. It could be language differences, as although the model is based on a multilingual BERT model, the nuances of the Czech language may not be well represented and may be important for the troll detection task. Additionally, the domain might be too widely missaligned, as the training data was collected from various platforms, various contexts and various time-frames. The missalignment of the training data could make it quite difficult for the model to use its learned knowledge on the specific Czech comments domain.

Recognizing the severe leimitations, we explored a strategy to overcome this problem by further fine-tuning the model on a small manually annotated dataset, which was menitoned previously when describing the annotation app.

The next step, described bellow, demonstrated more promising results.

5.3.2 Fine-Tuned Experiment

After seeing the poor performace of the "plug and play" approach, we decided to try to explore the fine tuning of the model through a few shot training approach. The goal was to see if the model could learn to adapt to the Czech language and domain by training on a small set of manually annotated users.

The model was trained for 5 epochs with a manually annotated dataset of 50 users. The performace of the the fine-tuned model on the Czech dataset improved significantly after fine-tuning, as can be seen in the new distribution of troliness scores in table below.

Statistic	Value
Mean	0.154568
Standard Deviation	0.253612
Median	0.001407
Maximum	0.999692

Table 5.7: Trolliness score distribution, fine-tuned model.

The results indicate improvement in the models ability to identify various levels of troll behavior. The mean trolliness score increased to 0.15, with a maximum score of 0.99. To better understand the model's behaviour, we visualized the distribution of trolliness predictions and scores across users in Figure 5.2.:

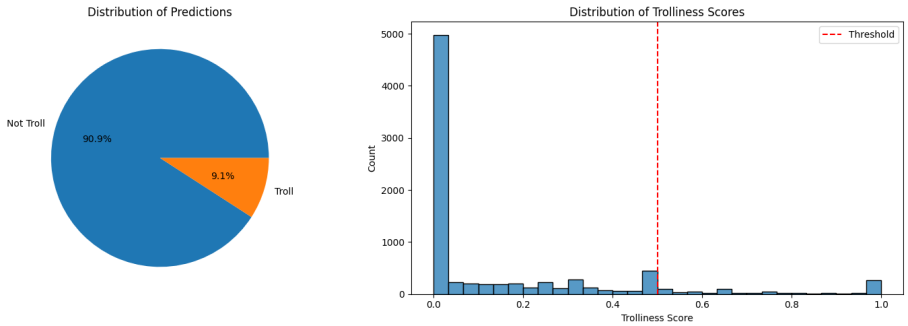


Figure 5.2: Trolliness score distribution, fine-tuned model.

The pie chart indicates that the model classifes only 9% of users as "Troll". This imbalance suggests that the model is conservative with assigning the a high trolliness score, which might not be a bad thing as it reduces the risk of false positives. And it also reflects what someone might expect, that only a small portion of users are trolls.

The histogram further illustrates the spread of trolliness scores. Most predictions cluster around zero and low troliness scores, however the model does still produce a range of scores reaching up to 1, showing that it does differentiate various degress of troll behaviour.

5.3.3 Manual Review of Predictions

To further understand the model's predictions, we manually reviewed a selection of users with high, low and uncertain trolliness scores. The goal was to see if the model's predictions aligned with our expectations and to identify any potential issues or biases in the model.

5.4 Adaptery Jsem Nakonec vynechal

Jazykovy adapter se neukazal ze by bohuzel pomahal a nasel jsem take i v literature zminene ze prave jen jazykovy adapter nestaci a mohl by vykon dokonce zhorsit. Tak asi nebudu psat nebo mam zminit experiment?

Chapter 6

Annotation and Evaluation

Chapter 7

Conclusion

Appendix A

Bibliography

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

[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.

[DBGJS21] Vlad Demsar, Jan Brace-Govan, Gavin Jack, and Sean Sands. The social phenomenon of trolling: understanding the discourse and social practices of online provocation. *Journal of Marketing Management*, 37:1–33, 03 2021.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[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.

[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].

[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.

[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.

[MPH21] Kristína Machová, Michal Porezaný, and Miroslava Hreškova. Algorithms of machine learning in recognition of trolls in online space. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000349–000354, 2021.

[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.

[RKR20] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how BERT works. *CoRR*, abs/2002.12327, 2020.

[SPN+24] Özgür Can Seçkin, Manita Pote, Alexander Nwala, Lake Yin, Luca Luceri, alessandro flammini, and Filippo Menczer. Labeled datasets for research on information operations, November 2024.

[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.

[TZL23] Lin Tian, Xiuzhen Zhang, and Jey Han Lau. Metatroll: Few-shot detection of state-sponsored trolls with transformer adapters. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1743–1753. ACM, April 2023.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[YZ23] Seyhmus Yilmaz and Sultan Zavrak. A context-sensitive word embedding approach for the detection of troll tweets, 2023.