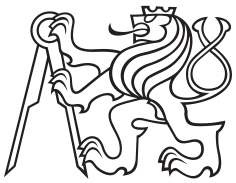
Bachelor thesis



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 online discussions create opportunities for connecting people and sharing information, they also come with diverse challenges like the spread of misinformation, polarization and disruptive behavior.

Given these challenges, 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.

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".

The term "trolling" 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.

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 as a giant or 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 closly related to the folklore definition. This shift in meaning reflects the evolution of online trolling from simple baiting to a more 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. 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]. 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 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 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 idenitfy harfmul contributions and contributors to online discussions and to explore possibilites for further research in thsi 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.

2.2 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.

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

The key innovation of Transformer models is the use of self-attention mechanisms. 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.

An 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.

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. In

this thesis, we will leverage pre-trained Transformer such models as BERT and try to fine tune them for the troll detection task.

Chapter 3

Related Methods

This chapter provides an overview of the methods and techniques used in the field of troll detection. It explores various approaches, including stylometry, sentiment analysis, and topic modeling, commonly employed to analyze online discussions and identify trolling behavior. Each section will discuss the principles behind these methods, their applications in troll detection, and their strengths and limitations.

3.1 Stylometry

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 ad-

vancements 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.

3.2 Sentiment Analysis

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 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.

3.3 Transformers

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.

Chapter 4

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.

4.1 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 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

4.2 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.

4.2.1 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]

4.2.2 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

[1] https://github.com/fivethirtyeight/russian-troll-tweets/

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.

4.2.3 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.

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

Chapter 5

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.

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.[RKR20]

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.

5.1 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.

5.2 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.

5.3 Training

The training of the model is done in two steps. First, the model is trained on the large labeled datasets collected from foreign domains. These include 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.

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].

[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.

[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.

[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.

[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.