Identification of Critical Sub-Components in Conversations

Dhruva ManishKumar Patel
Arizona State University
Tempe, Arizona
dpate266@asu.edu

Priyesha Jethi
Arizona State University
Tempe, Arizona
pjethi@asu.edu

Sagar Rajesh Kumar Sinha
Arizona State University
Tempe, Arizona
ssinha78@asu.edu

Swati Kashyap
Arizona State University
Tempe, Arizona
skashy10@asu.edu

ABSTRACT

Online discussions encompass a spectrum of discourse modes, ranging from collaborative rational exchanges to adversarial interactions marked by disruptive elements. This project delves into the intricate dynamics of these modes by deciphering latent components such as hate speech, trolling, and deception within Reddit conversations. Through a multifaceted linguistic analysis, the study aims to unravel the interplay between various discourse elements, providing insights into their overall impact. Furthermore, by coalescing the feature exploration with structural properties derived from graph neural networks, the research endeavors to model the dynamics of online discussions more comprehensively. The findings of this study hold the potential to bridge ideas from seemingly independent domains, fostering a more holistic understanding of online discourse dynamics and their implications.

1 INTRODUCTION

Since their inception, social media platforms have garnered immense popularity. In recent years, these forums have witnessed a surge in user adoption rates across various strata of society. Nonetheless, almost all of these platforms grapple with one or more of the critical sub-problems we aim to study, which can have a significant psychological impact on individuals. The prevalent issues of biased ideologies, the development of racial and sexist stereotypes, religious polarization, and others can be attributed to the prevalence of the tasks we aim to capture.

Among the most prevalent and pernicious sub-components plaguing social media platforms is hate speech. It has been discussed before that there exists hate speech on online platforms. This insidious form of expression directly contributes to the exacerbation of polarization across diverse aspects of our social fabric, fostering an environment of intolerance, discrimination, and division. Consequently, a significant portion of academic research in the domain of this discourse detection has been dedicated to studying and mitigating the pervasive issues of hate speech and trolling, which involves the deliberate act of provoking, disrupting, or antagonizing others through inflammatory, deceptive, or off-topic messaging.

Trolling, on the other hand, encompasses a range of behaviors with both positive and negative consequences. The positive aspect of it can be playful, and foster creativity and critical thinking, while negative trolling can be disruptive and deliberately inflammatory. In NLP, this creates a further hurdle, as humor and irony are notoriously difficult for algorithms to detect, making it challenging to differentiate between playful and malicious trolling.

Deception is widely prevalent in editorials, op-eds published by news agencies, journalists, and other individuals or collective groups. Numerous instances exist where manipulated content is shared online to suit a particular political or ideological narrative. Deception serves as a major source of misinformation, and its prevalence can be particularly worrisome during critical times, such as the recently occured COVID-19 pandemic. The dissemination of deceptive information can have far-reaching consequences, exacerbating confusion, undermining trust in authoritative sources, and potentially leading to harmful decisions or actions.

Although they have been well-defined in literature, most of the existing studies regarding these sub-components are really specific and fixated to the problem being solved at hand, ignoring other components that might be actually useful for a score boost. This causes the loss of generalizability, when the problem is seen in the light of other related discourse elements. For example, if the researchers are dealing with a complex problem such as that of modelling the cancel culture, then it might help to view the problem along a greater breadth as well. Cancel culture or call-out culture is a modern form of ostracism in which someone is thrust out of social or professional circles – whether it be online, on social media, or in person.

The reason or the cause of this phenomenon "Cancel culture" [6], [1] prevalent in the current social media can be attributed to many other smaller sub-problems. One might see this problem from the perspective of failure in the moderation of the hate speech over the platform. The fact that the discussions online are not coherent and in sync with each other, might lead to confusion followed by misinterpretation of ideas. This could lead to heated conversations. It is important to note that people online are watching and even participating in this hot mess of a conversation that unravels itself. It is natural for people to start taking sides.This would lead to some degree of polarization (yet another component which we haven't considered in our scenario) in the conversation based on whether people agree or disagree with the opinion under observation.

Thus we have some idea of how these various sub-components could affect the cadence of the conversation. As a part of this study, we are trying to aggregate the insights gained from previous attended problems to enhance the formulation framework for future research problems. Precisely, we can define two clear objectives which pave the way of our endeavour going ahead.

Objective 1: *Can we perform a multifaceted linguistic analysis on Reddit conversations which maps the interplay between various discourse elements mentioned above?*
Objective 2: *Can we utilize the feature exploration from the first objective, and coalesce them with structural properties derived from graph neural networks to shape the dynamics of online discussions?*

2 RELATED WORK

The current literature on identifying critical sub-components in online conversations encompasses a spectrum of methodologies, each contributing uniquely to the understanding and analysis of complex discourse elements. Key studies such as those by [7], [8], and [10], have explored automated interventions in hate speech and the intricacies involved in detecting hate speech on social platforms. The effectiveness of Natural Language Processing (NLP) techniques, including Seq2Seq, VAE, and Reinforcement Learning, was investigated in generating counter-narratives to hate speech, providing a comprehensive evaluation of the models based on both automated metrics and human assessment.

These works also address the limitations in the scope of data collection, often confined to particular platforms and temporal ranges, potentially affecting the findings' applicability and generalizability. For instance, [7] focused solely on datasets from Reddit and Gab, while [8] extended the study to include 4chan and 8chan, assessing the prevalence and nature of hate speech across these platforms.

In the context of trolling, [5] proposed a nuanced categorization scheme that includes the troll's intention and the responders' interpretation and strategy, alongside investigating linguistic features and word embeddings as potential feature sets for trolling prediction.

Regarding deception, [9] highlighted the absence of a universal set of verbal cues for deception detection, emphasizing the need for domain-specific approaches and the challenges in identifying subtle deceptive cues.

Additionally, studies like those by [2] and [4] have advanced the field of feature learning for network structures, offering scalable algorithms and combining semantic understanding with structural modeling. Node2Vec's biased random walks, for instance, have shown promise in analyzing the diverse interactions inherent in online conversations, while BertGCN's integration of BERT with GCN presents an innovative method for transductive text classification.

The aforementioned studies form a bedrock for future research, especially in applying and refining Graph Neural Networks (GNNs) for the structural analysis of discourse elements such as hate speech, trolls, and deception. These methodologies highlight the strengths and potential of various computational approaches to tackle the multifaceted nature of online discussions and offer avenues for future work to bridge gaps in data collection, analysis, and the interpretation of complex social interactions.

3 FORMAL DEFINITION OF TASK

To define the task, firstly we should formally introduce all the elements that constitute a canonical conversation over the online social media platforms. The study or the framework that we are preparing can be easily extended to have multiple data sources. Let the set of Data sources be $D = \{D_1, D_2, D_3, ...\}$. For our testbed we are only dealing with Reddit - D_1.

Every source D_i is an unordered collection of many conversations/submissions/threads c_i such that $D_i = \{c_1^{'}, c_2^{'}, c_3^{'}....c_N^{'}\}$ where N is the total number of conversations from that source and can vary from source to source. Data from all the D_i are then combined and segregated on the basis of their relevance with respect to the set of selected Sub-Problems $S \in \{S_1, S_2, S_3, ..., S_n\}$. Once separated based on the tests for each sub-problem, each sub problem S_i gets its own data, which is yet another unordered set of conversations c_i from the combined data i.e. $D_{S_i} = \{c_1, c_2, c_3..., c_n\}$. It is important to note that a particular c_i can belong to one or all of the sub-components $S \in \{S_1, S_2, S_3, \ldots, Sn\}$.

Let's break down the process: We'll focus on one sub-component at a time, such as S_1. Each sub-component will have associated data, denoted by D_{si}. Within a sub-component, conversations are represented by c_i. These conversations are ordered collections of responses (replies or utterances) u_i to a post p_i. In other words, each conversation c_i is a set containing the post p_i and an ordered list of utterances $\{u_1, u_2, u_3, \ldots, u_n\}$, where n is the total number of utterances in the conversation.

Each of the c_i are converted into a graphical representation $G_{c_i} = (V, E, V_f, E_f)$ where V are the nodes, E stands for the edge relations between $\{V_1, V_2, ..., V_x\} \in V$. Each node and edge can have a set of features associated with it and they are represented using the Node feature matrix V_f and E_f respectively.

4 TESTBED DEVELOPMENT

One of the major outcomes of this study would be a research testbed that can be used for further work in the related areas. This test-bed could potentially facilitate researchers from related fields to exemplify their works on conversational behaviors. There are a plethora of options when it comes to these sub-components, but many of them don't really capture the essence of the conversation because they constitute stand-alone sentences. In order for us to tap into this latent information, we need the entire conversation which includes "comment trees".

4.1 Platform Choice

Reddit, a unique social media platform, can be described as a vast collection of tightly-knit communities centered around shared interests. These communities, known as "subreddits," act as sub-groups within the larger platform. Each subreddit thrives on user-generated content, with members creating "submissions" (posts in Reddit's API terminology) that can include comments in a nested reply structure, mirroring the common interface of most social media platforms.

We utilized this platform since it fosters a unique environment for conversation compared to other platforms. For one, it utilizes a system of subreddit rules with automated checks by bots alongside human moderators for each community. This creates a space that's generally considered self-moderated by the users themselves. Participants actively downvote content that violates the rules, creating a strong incentive for adherence. This way, the platform enables us to cherry pick subreddits corresponding to different subcomponents since all the participants in a particular community

put forth their views in a manner consistent with the subreddit's ideology or knowledge-sharing mechanisms.

Previously, extracting data from Reddit relied heavily on the official Reddit API or third-party tools like PRAW and PushShift. However, policy changes by Reddit made bulk data extraction challenging. To overcome this hurdle, we explored alternative solutions and discovered platforms like Convokit mark it, which provide access to vast, pre-stored repositories of Reddit data hosted on dedicated databases. This approach proved ideal for our specific use case.

4.2 Annotation Scheme

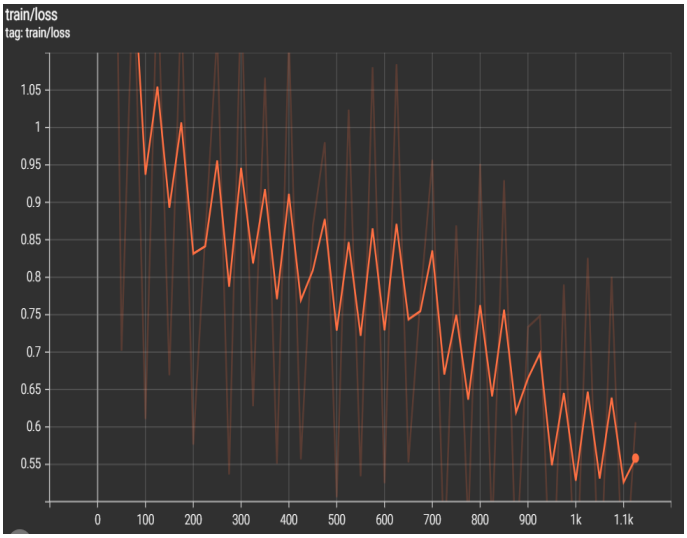

Figure 1: Fine-tuning loss curve for Llama 7B on Ethos hate speech dataset, showing oscillatory decrease over training steps

The annotation scheme is built with an aim to decipher node-edge relationships wihth an unlabelled dataset of conversational data. As a first step, we tried to finetune Llama-2-7b chat model for sub-component classification, but due to computational constraints and large model inference times, we have planned to use a more primitive language model such as Bert or Roberta. Through this, we aim to achieve swift inferencing capabilities and achieve the requisite labels.

To ensure the reliability of our data labeling process, we will be utilizing a specialized inhouse annotation application where a small subset of conversations will be meticulosuly reviewed and labeled in turns by a trio of annotators. The integrity of our annotations will be rigorously checked against inter-annotator agreement matrices, allowing us to establish a consensus and maintain consistency within the annotations. This comprehensive validation scheme would be critical for reinforcing the accuracy of our conversational analysis.

4.3 Data Analysis

In this section of the paper, we will target **Objective 1** from above. Our preprocessing pipeline focuses on comments within active conversations on Reddit. We start by loading the conversations dataframe. Here, we filter out comments with zero replies, ensuring we analyze comments that have generated discussion. Next, we clean the utterances associated with these comments. This cleaning step removes various Reddit-specific formatting elements, including newline characters, quotes, bullet points, links, dates, strikethroughs, spoilers, hashtags, and emojis. This standardization ensures the text is consistent for further analysis. We are also planning to implement an additional filtering step. This recursive filter aims to identify and remove comments from known bots, deleted speakers, and their associated replies. This will be done while preserving the original threaded structure of the conversations. To achieve this, we have already integrated a level-order Depth-First-Search algorithm in the preprocessing pipeline. Post that, we analyzed our dataset(s) for the presence of each of the sub-components.

4.3.1 Hate Speech. For extracting discriminative hate speech features, I have utilized sentiment analyzer by [3], which is a well-regarded tool adept at gauging polarity in the context of social media. Complementing this, we also utilized Perspective API, a tool for the assessment and mitigation of toxicity within online discourse. Some results of our analysis are as follows:

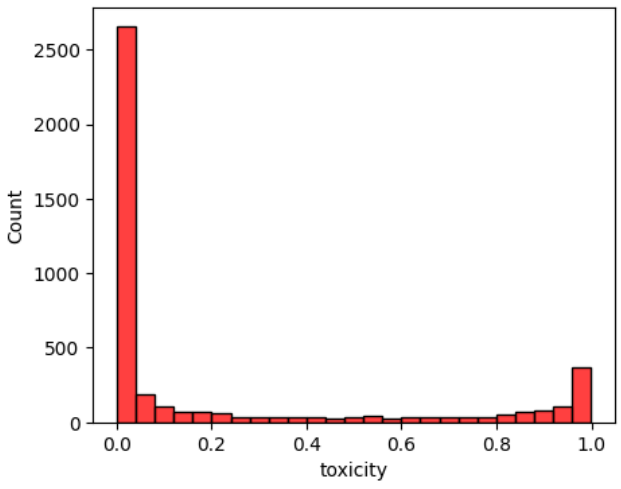

Figure 2: Histogram of toxicity scores across the dataset. The distribution is predominantly concentrated at the lower end, with the majority of comments registering minimal toxicity. Occasional spikes towards the higher end of the toxicity scale indicate the presence of outliers or less frequent highly toxic comments, characterizing a right-skewed distribution.

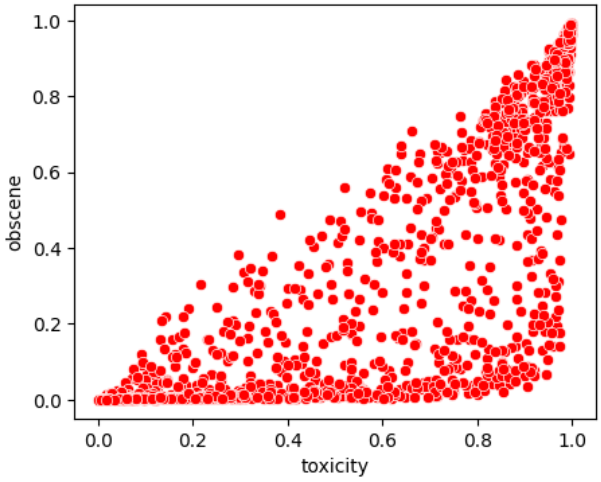

Figure 3: Scatter plot displaying the positive correlation between 'toxicity' and 'obscene' ratings in online content analysis. Data points are more densely populated towards the higher end of both axes, indicating a trend where instances rated highly toxic are often also rated as highly obscene. This suggests a strong association where higher toxicity levels in the content are likely to be accompanied by higher levels of obscenity.

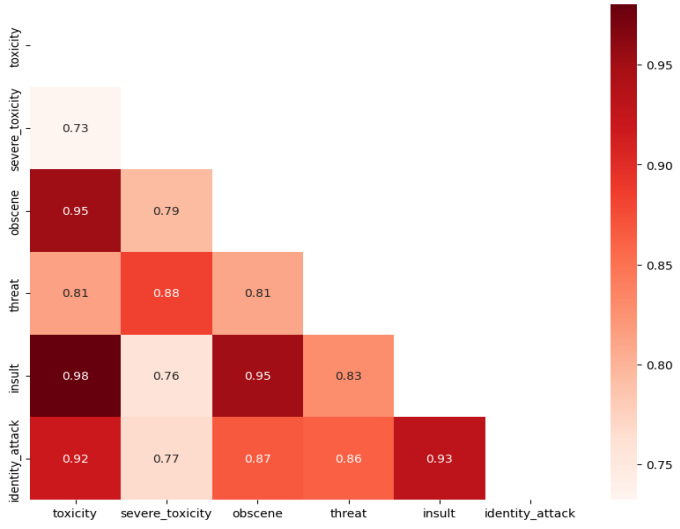

Figure 4: Heatmap of correlation coefficients between various toxicity dimensions in textual data. Strong positive correlations are depicted by darker reds, with the highest correlation between 'toxicity' and 'insult' (0.98), suggesting a significant overlap in these two dimensions. The heatmap underscores the interconnected nature of different forms of toxic expression in online communication.

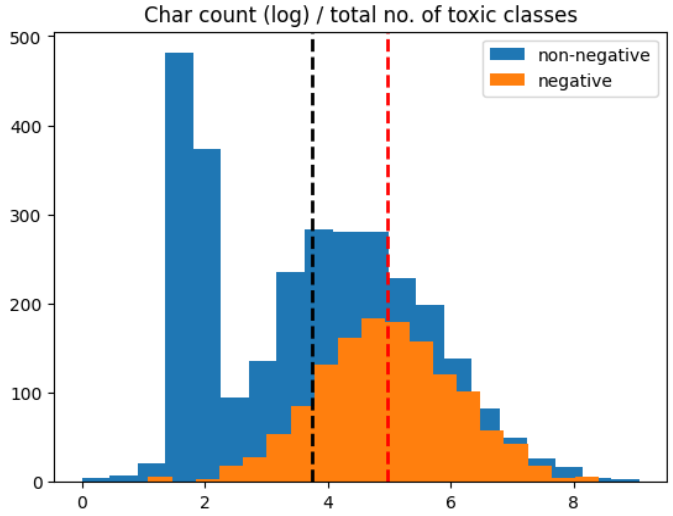

Figure 5: Histogram comparing the distribution of character counts on a logarithmic scale between 'clean' and 'dirty' text classes. The 'clean' class (blue) shows a higher frequency of texts with shorter character counts, peaking at lower values, while the 'dirty' class (orange) exhibits a broader distribution, indicating a greater variability in text length with toxic content. The dashed lines may represent the mean or median, highlighting the text length differences between the categories.

4.3.2 Trolls. In this section, we present the methodology and results of our data analysis aimed at identifying critical sub-components in conversations within the Formula 1 subreddit, specifically focusing on the detection of trolling behavior.

(1) **Data Cleaning** The first step in our analysis involved cleaning the dataset to remove irrelevant elements such as emojis, links, strike-throughs, and bullet points. This process ensures that the data is standardized and conducive to further analysis.
(2) **Word Lists Creation** Two distinct lists of words were created to facilitate the identification of trolling behavior:
   (a) List 1: Variants of 'Troll' and related words * This list includes all forms of the words 'troll' and related words, such as 'trolling', 'trolled', 'rude', 'rudeness', etc.
   (b) List 2: Highly Offensive Words and Phrases * This list encompasses highly offensive words, swear words, and short phrases commonly associated with trolling behavior.
(3) **Feature Assignment** Each comment in the dataset was evaluated based on the presence of words from the created lists. The following criteria were employed for feature assignment:
   (a) Binary Feature Assignment: Comments were assigned a binary feature, denoted as 1 if they met either of the following conditions: The comment contains child comments or direct replies with any form of the word 'troll'. The comment contains one or more words/phrases

from the second list of offensive words. Comments not meeting these criteria were assigned a binary feature value of 0.

(4) **Classification and Data Visualization** Following feature assignment, a classification analysis will be performed to distinguish between trolling comments and normal comments. The dataset was divided into two classes: trolling comments (1) and normal comments (0). Various classification algorithms such as logistic regression, support vector machines, or decision trees could be employed for this task. The data can be visualized using plots to illustrate the percentage of trolling comments versus normal comments. This visualization aids in understanding the prevalence of trolling behavior within the Formula 1 subreddit and allows for comparisons between different time periods or specific threads.

Through the described methodology, we aim to effectively identify critical sub-components in conversations within the Formula 1 subreddit, specifically focusing on the detection of trolling behavior. The classification and visualization techniques to be utilized will provide insights into the prevalence and distribution of trolling comments, thereby enabling a better understanding of community dynamics and facilitating moderation efforts.

4.3.3 Deception. In this section, we introduce a study focused on detecting deception in a broad range of contexts, along with an examination of deceptive patterns related to gender and age. We are currently in the process of developing deception classifiers using the SVM algorithm along with different sets of features. As part of our ongoing work, we are conducting five-fold cross-validation, where we train on 80% of the users each time and assess performance on the remaining 20%İn order to assess how the volume of data influences the learning process of the classifier, we generated learning curves for various feature sets by gradually increasing the amount of data. We conducted evaluations using five-fold cross-validation at each incremental data fraction.

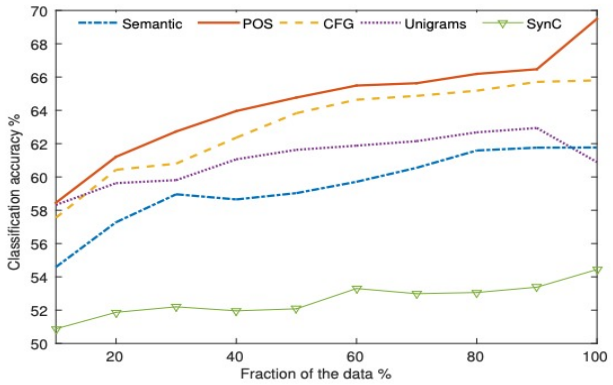

Figure 6: Learning curves for deception detection using five feature sets

(1) **Dataset Collection** We assembled a novel open domain deception dataset featuring one-sentence truths and lies, along with demographic particulars such as gender and age of the contributors.

(2) **Feature Extraction** Extracting various sets of features from the dataset involved:
   (a) Unigrams: Employing a basic bag-of-words technique to represent the text.
   (b) Shallow and Deep Syntax Features: Deriving part of speech tags and lexicalized production rules from Probabilistic Context-Free Grammar (PCFG) trees.
   (c) Semantic Features: Utilizing the LIWC lexicon, which encompasses 80 semantic classes, to quantify the number of words in a sentence belonging to specific semantic categories.
   (d) Readability and Syntactic Complexity Features: Incorporating readability scores (e.g., Flesch-Kincaid, Gunning Fog) and syntactic complexity measures derived from syntactic analysis of each sentence.

(3) **Classifier Training and Evaluation** Trained Support Vector Machine (SVM) classifiers using different feature sets, and assessed their performance using a five-fold cross-validation strategy, wherein training was conducted on 80% of the data and testing on the remaining 20%

(4) **Analysis of Language Differences** Analyzing linguistic disparities in deceptive content based on the gender and age of the deceivers involved employing the Linguistic Inquiry and Word Count (LIWC) lexicon to identify dominant semantic word classes associated with truths and lies provided by males and females, as well as differences across different age groups.

5 BASELINE IMPLEMENTATION

In this section we will implement discuss conversion of raw data to graph format.

5.1 Model Architecture

The model will be a graph attention model which would be modelled upon forests of different sub-component data, and each of the forests would contain multiple conversation graphs. For testing, a conversation wuld be compared with all the discourse embeddings and then an attention or co-occurence score would be generated highlighting the interplay of different problems within each comment.

5.2 Metric Evaluation

TBA. In this section I will evaluate the performance of my model using accuracy, precision, recall, F1 score, etc.

5.3 Discussion

TBA. The discussion would proceed depending upon the performance of our model on testing data. Upon that, we would also devise a future course of action.

6 CONCLUSION

Our study, pivoting on the structural and linguistic nuances of Reddit dialogues, would establish a data-driven approach to discern critical sub-components like hate speech, trolling, and deception. We aim to meticulously classify individual comments using advanced graph neural networks and affirm our annotation accuracy through a custom-built application, ensuring robust inter-annotator reliability. This convergence of computational techniques and human oversight presents a promising direction for future research, aiming to expand our analysis across diverse social media platforms and further refine the computational models to encapsulate the evolving dynamics of online interactions. Our findings lay the groundwork for enhanced moderation tools and a deeper comprehension of discourse on digital platforms.

REFERENCES

[1] Meredith D. Clark. 2020. DRAG THEM: A brief etymology of so-called "cancel culture". *Communication and the Public* 5, 3-4 (2020), 88–92.

[2] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.

[3] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225.

[4] Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. Bertgcn: Transductive text classification by combining gcn and bert. *arXiv preprint arXiv:2105.05727* (2021).

[5] Luis Gerardo Mojica. 2016. Modeling trolling in social media conversations. *arXiv preprint arXiv:1612.05310* (2016).

[6] Thomas S Mueller. 2021. Blame, then shame? Psychological predictors in cancel culture behavior. *The Social Science Journal* (2021), 1–14.

[7] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. arXiv:1909.04251 [cs.CL]

[8] Diana Rieger, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and Georg Groh. 2021. Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media+ Society* 7, 4 (2021), 20563051211052906.

[9] Anna Vartapetiance and Lee Gillam. 2014. Deception detection: dependable or defective? *Social Network Analysis and Mining* 4 (2014), 1–14.

[10] Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web* 10, 5 (2019), 925–945.