

# Identification of Critical Sub-Components in Conversations

Dhruva ManishKumar Patel

Arizona State University  
Tempe, Arizona  
dpate266@asu.edu

Sagar Rajesh Kumar Sinha

Arizona State University  
Tempe, Arizona  
ssinha78@asu.edu

Priyasha Jethi

Arizona State University  
Tempe, Arizona  
pjethi@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 occurred 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” Mueller[8], D. Clark[2] 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 online conversations which maps the interplay between various discourse elements mentioned above?*

**Objective 2:** *Can we utilize the feature exploration from the first objective, and study them concurrently 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 Qian et al.[9], Rieger et al.[10], and Zhang and Luo[12], 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, Qian et al.[9] focused on datasets from both Reddit and Gab, while Rieger et al.[10] extended the study to include 4chan and 8chan, assessing the prevalence and nature of hate speech across these platforms.

In the context of trolling, Mojica[7] 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, Vartapetian and Gillam[11] 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 Grover and Leskovec[3] and Lin et al.[6] 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  $D = \{D_1, D_2, D_3, \dots\}$ . 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  $D_i = \{c'_1, c'_2, c'_3, \dots, c'_N\}$  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  $S \in \{S_1, S_2, S_3, \dots, S_n\}$ . Once separated based on the tests for each sub-problem, each sub problem  $S_j$  gets its own data, which is yet another unordered set of conversations  $c_i$  from the combined data i.e.  $DS_j = \{c_1, c_2, c_3, \dots, c_n\}$ . It is important to note that a particular  $c_i$  can belong to one or all of the sub-components  $S \in \{S_1, S_2, S_3, \dots, S_n\}$ .

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  $DS_i$ . 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  $\{u_1, u_2, u_3, \dots, u_n\}$ , 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  $\{V_1, V_2, \dots, V_x\} \in V$ . 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 DATA 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 sub-components 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 PSAW. However, policy changes by Reddit made bulk data extraction challenging. To overcome this hurdle, we explored alternative solutions and discovered platforms like the one provided by Chang et al.[1], 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: Overview of sub-problem annotation platform

The annotation scheme is built with an aim to decipher node-edge relationships within 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 decided to use a more primitive Bert or Roberta-based task-specific finetuned models. Through this, we leveraged swift inferencing capabilities of these models to get faster annotations.

First of all, we utilized a specialized inhouse annotation application where a small subset of conversations were meticulously reviewed and labeled in turns by a trio of annotators. The integrity of our annotations were rigorously checked against inter-annotator agreement matrices, allowing us to establish a consensus and maintain consistency within the annotations. This comprehensive validation scheme was critical for reinforcing the accuracy of our conversational analysis.

Thereafter, we employed a Bert-based pre-trained model detoxify, which is trained on a substantial dataset of toxic Wikipedia

comments, and performed few-shot classification on the annotated dataset to expand the labels for hate speech for the entirety of training data.

For the purpose of annotating instances of trolling within our dataset, we implemented a heuristic methodology, since we couldn’t find a good pretrained models for our implementation scenario and also due to time constraints. This approach involved a binary classification of the dataset, wherein we devised two distinct lists of keywords. Firstly, we curated a compilation of terms related to the concept of “trolling”. Any instance of an utterance containing one of these predetermined keywords resulted in the immediate parent utterance being categorized as indicative of trolling (marked as 1), while those devoid of such terms were designated as non-trolling (marked as 0).

Secondly, we assembled a catalog of highly provocative words and phrases for our chosen subreddit(s). This compilation encompassed terms frequently employed for the purpose of trolls. Should any of these designated keywords manifest within an utterance, it was consequently flagged as indicative of trolling.

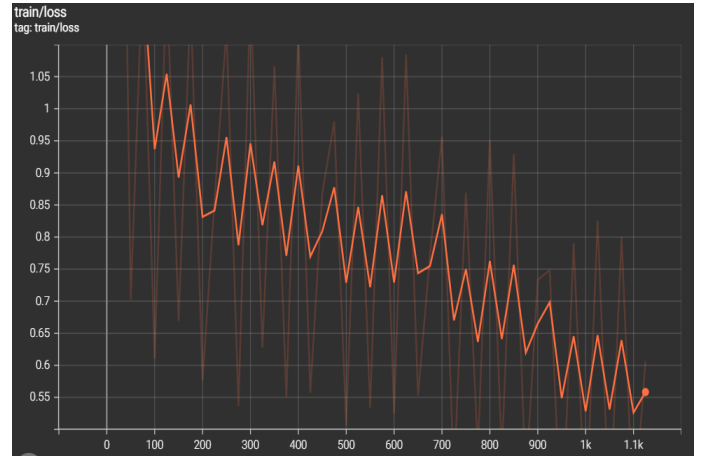


Figure 2: Fine-tuning loss curve for Llama 7B on Ethos hate speech dataset, showing oscillatory decrease over training steps

## 4.3 Preliminary Annotation Results

Two subreddits, **r/The\_Donald** and **r/PussyPass**, were chosen for our annotation scheme. A sample of 50 conversations was extracted from each subreddit, focusing on conversations with a comment range of 5-10. Conversations shorter than 5 comments were excluded since we believe that they lacked sufficient context for effective learning by Graph Neural Networks (GNNs). Lengthy conversations (more than 10 comments) were also excluded to maintain a conservative approach and maintain a good annotation quality. The implemented annotation scheme achieved an average Fleiss’ Kappa agreement score of 0.67 (0.65 for the first corpus and 0.69 on the second corpus). The annotation process adhered to the following guidelines:

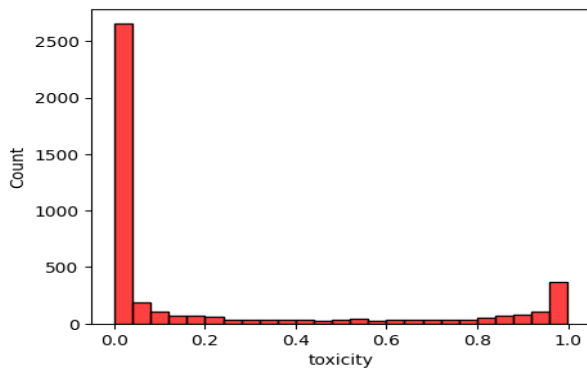
Our labeling process prioritizes context. Deleted content (indicated by “[removed]”, etc.) automatically receives a 0 label. Parent

comments without children are also labeled 0, as they lack the context provided by a child’s response. When only a child comment exists, we make an informed decision based on its standalone content. If both parent and child are present, we evaluate whether the child relates to the parent’s themes or introduces problematic elements on its own, determining the labels accordingly.

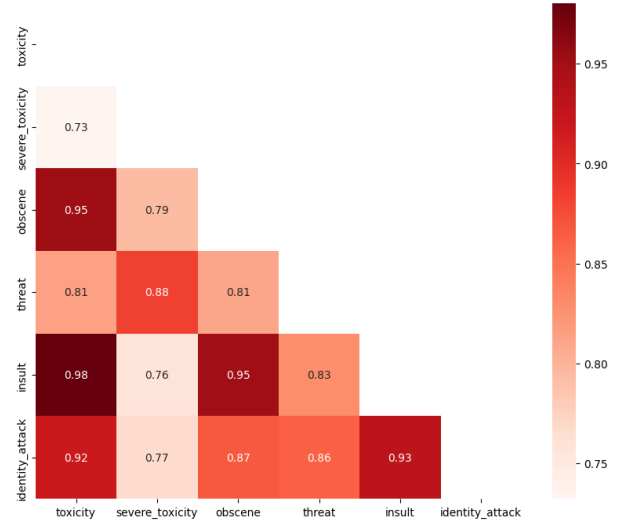
#### 4.4 Data Analysis

In this section of the paper, we will target **Objective 1** from above. Our preprocessing pipeline focuses on comments within active conversations on Reddit. Here, we have handpicked data from Convokit for two banned subreddits, namely **r/The\_Donald** and **r/PussyPass**. To begin with, 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 new-line characters, quotes, bullet points, links, dates, strikethroughs, spoilers, hashtags, and emojis. This standardization ensures the text is consistent for further analysis. Next, we used a recursive filter aims to identify and remove comments from known bots, deleted speakers, and their associated replies. This was 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.

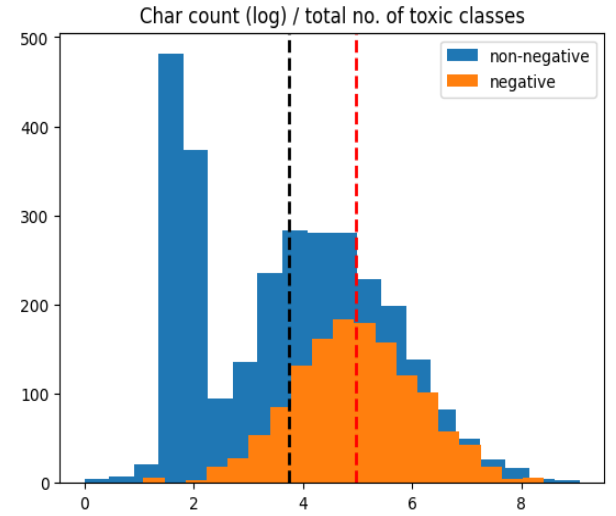
**4.4.1 Hate Speech.** For extracting discriminative hate speech features, we used a sentiment analyzer by *Hutto and Gilbert*[4], which is a well-regarded tool adept at gauging polarity in the context of social media. Complementing this, we also utilized Perspective API, a tool for the assessment and mitigation of toxicity within online discourse. Some results of our analysis are as follows:



**Figure 3:** 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 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.



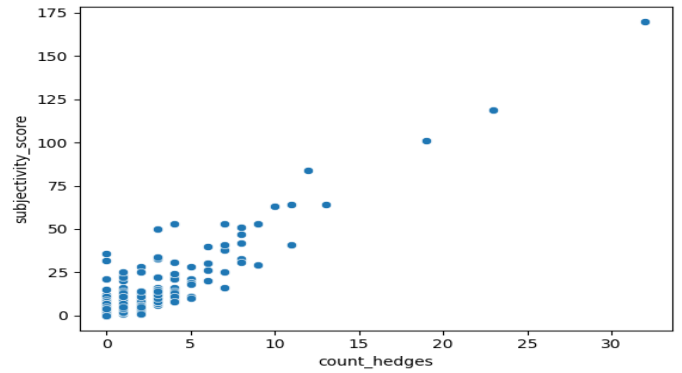
**Figure 5:** Histogram comparing the distribution of character counts on a logarithmic scale between ‘clean’ and ‘dirty’ text classes. The ‘clean’ class (blue) shows a higher frequency of texts with shorter character counts, peaking at lower values, while the ‘dirty’ class (orange) exhibits a broader 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.

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

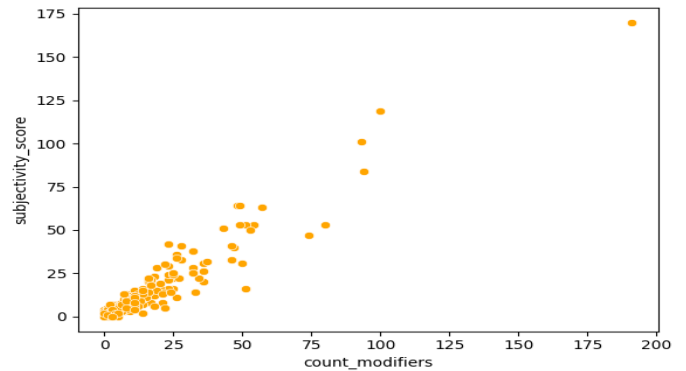
- (1) **Data Cleaning** The first step in our analysis involved cleaning the dataset to remove irrelevant elements such as emojis, links, strike-throughs, and bullet points. This process ensures that the data is standardized and conducive to further analysis.
- (2) **Word Lists Creation** Two distinct 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 specific to the topic of subreddit, i.e. Donald Trump. This list contains phrases associated with Donald Trump but used in a negative view.
- (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 was 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 effectively identified critical sub-components in conversations within the Donald subreddit, specifically focusing on the detection of trolling behavior. The classification techniques utilized will provide insights into the prevalence and distribution of trolling comments, thereby enabling a better understanding of community dynamics and facilitating moderation efforts.

**4.4.3 Deception.** Previous research has shown that detecting deception involves analyzing various cues, including verbal, non-verbal, and vocal indicators. In our approach, we used functions to measure the frequency of adjectives, adverbs, modal verbs, references to groups, and the subjectivity score on the processed non-verbal/text data. Here are some key findings from our analysis:



**Figure 6:** This scatter plot shows the intricate interplay between subjectivity score(proxy for personal opinion within textual communication) and hedge count (frequency of modal verbs indicating uncertainty). Most data points exhibit low subjectivity scores, potentially indicating deceptive intent, alongside low hedge counts, suggesting confidence and certainty in the speaker's utterance. This convergence suggests a prevalence of linguistic ambiguity and uncertainty within the dataset, which may signify the presence of deceptive elements.



**Figure 7:** This scatter plot explores the complexities of deception detection by scrutinizing the correlation between subjectivity score (representing personal opinion in textual communication) and count modifier (reflecting the frequency of adjectives and adverbs). The majority of data points showcase low subjectivity scores, suggesting a possible inclination towards deceptive intent. Additionally, the prevalence of low count modifier values implies uncertainty in the user's claims, further suggesting potential indicators of deception.

The exploration and analysis of features claims potential existence of detecting in utterance. Through this analysis, it was found that distinguishing deception from other phenomena like trolling and spreading false information is challenging without considering non-verbal and vocal cues. Factors like anonymity, linguistic variability, frequent use of sarcasm and irony, and data quality

pose significant challenges in analyzing non-verbal cues, making deception detection complex. Additionally, the lack of a definitive pre-established model for annotating deception detection led to its exclusion from our model development.

## 5 BASELINE IMPLEMENTATION

In this work, we focus on the task of node classification in conversation graphs, which is our **Objective 2** from above. Given a conversation graph, the task aims to classify whether a comment (represented as a node) contains hate speech or troll based on the context of the neighbouring nodes (other comments) to which it is directly/indirectly connected.

### 5.1 Model Architecture

We have implemented a custom Graph Convolutional-based architecture, inspired by Kipf and Welling[5], utilizing a structured approach to node classification and employing specialized components to optimize performance and accuracy. The architecture comprises two primary modules: the Convolutional Representation for Discrimination (CRD) for feature enhancement and the Classifier (CLS) for mapping enhanced features to output classes. Using PyGeometric's "GCNConv" module, the model efficiently processes graph-structured data, incorporating strategies like dropout and ReLU activation to manage overfitting and non-linearities effectively.

Specific configurations of the model are as follows:

- **CRD Module:** Maps input features to 16 hidden units using GCNConv, includes ReLU activation and dropout ( $p=0.5$ )
- **CLS Module:** Maps 16 hidden units to output classes using GCNConv; applies log softmax for output normalization
- **Optimizer:** Adam optimizer with settings for learning rate (0.003) and weight decay (0.0005)
- **Training:** Configured to train for upto 100 epochs across 5 runs with early stopping based on validation performance
- **Dropout rate:** Set at 0.5 in CRD module to prevent overfitting
- **Activation:** ReLU used in CRD module to introduce non-linearity

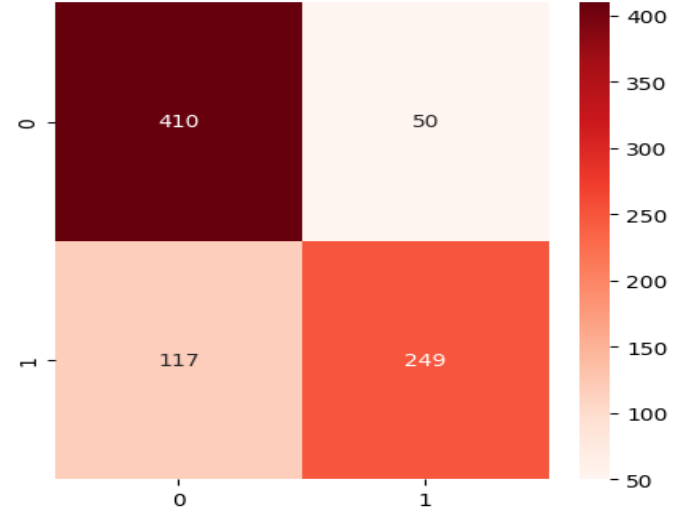
## 6 RESULTS

In the investigation of online discourse dynamics utilizing custom Graph Convolutional Network (GCN) model, we get the following preliminary results.

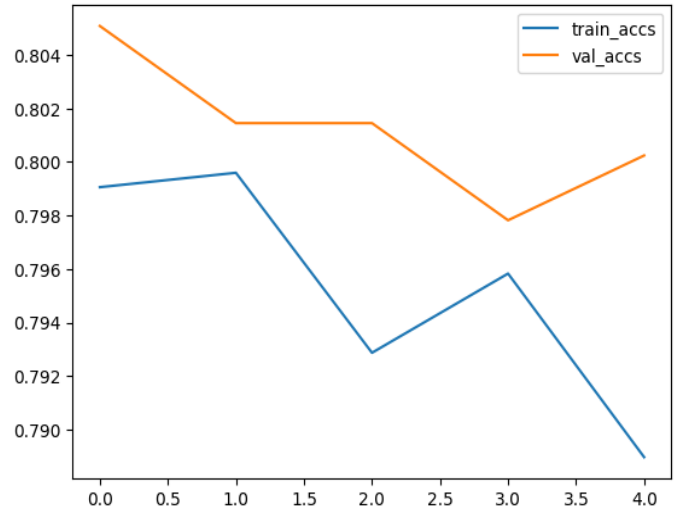
**6.0.1 Hate Speech.** The classification results of our base model on hate speech are described ahead in the table and plots as follows:

**Table 1: Hate Speech Classification Metrics**

Label	Precision	Recall	F1-Score	Support
0	0.78	0.89	0.83	460
1	0.83	0.68	0.68	366



**Figure 8:** The given confusion matrix shows the model proficiency in correctly identifying non-hate speech, but also highlights its comparative struggle with correctly classifying instances of hate speech against false negatives. The disparity suggests a need to improve the model's sensitivity to the less frequent hate speech occurrences, a common challenge in imbalanced datasets



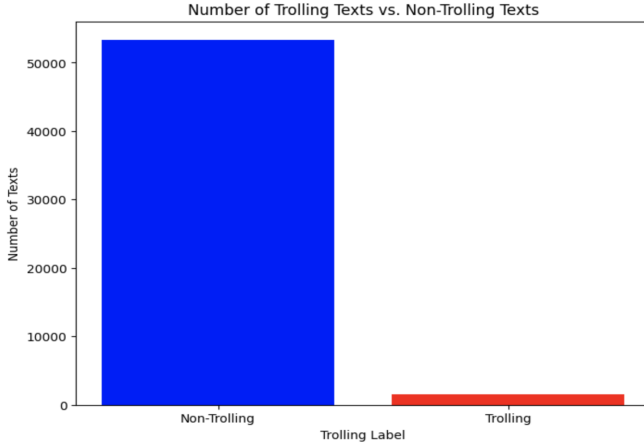
**Figure 9:** Accuracy trends for a GCN model on hate speech and troll detection show training challenges, with overfitting indicated by higher training accuracy and a notable dip suggesting model instability. Validation accuracy remains stable but plateaus, highlighting the need for refined training approaches to improve model generalization

**6.0.2 Trolls.** The classification results of our base model on trolls are described ahead in the table and plots as follows:



**Table 2: Troll Classification Metrics**

Label	Precision	Recall	F1-Score	Support
0	0.72	0.89	0.80	460
1	0.80	0.67	0.72	366

**Figure 10: Number of Trolling Text vs Non-Trolling Texts**

## 7 DISCUSSION

The ability of GCN’s to integrate linguistic features with structural network properties shows a good potential in identifying subtle nuances that can differentiate between different types of discourse. The analysis of our GCN model’s performance revealed it to be effective in learning the dataset’s patterns, as evidenced by the decreasing (though not consistent) training loss. However, fluctuations in validation loss and the observed discrepancy in accurately classifying the minority class indicates the necessity for more careful monitoring of model complexity and adequate hyperparameter tuning.

Comparatively, the GCN model addresses some of the limitations observed in conventional text-based classification systems. By incorporating structural features of conversation threads into the analysis, our approach offers a more holistic understanding of discourse patterns, which is often missing in models that treat text data as isolated instances. This structural analysis capability distinguishes our methodology from others, as highlighted in studies by Rieger et al.[10], which predominantly focusses on content-based analysis without considering the relational dynamics among conversation components. Additionally, Qian et al.[9] employed traditional NLP methods to detect patterns within text data, whereas our work integrates both linguistic features and structural relationships within conversation threads using GCNs. This dual focus allows for a more nuanced detection mechanism, as the graph-based model can potentially capture not only the content but also the interactions between different parts of the conversation., which we believe is crucial for understanding the broader discourse context.

Despite its potential advantages, the study faces limitations, particularly concerning the dataset’s scope and diversity. The reliance

on data from predominantly polarized subreddits may introduce bias into the model, potentially limiting its generalizability to less contentious or more diverse online environments. The interpretation of irony and sarcasm in trolls remains a significant challenge, as these require not only linguistic cues but also an understanding of the speaker’s intent and contextual clues that are often not evident in the text alone. Furthermore, similarities between lying and deception makes differentiation challenging. Lies are clear cases of stating something known to be false. While, deception involves a range of behaviors like omitting information, saying half-truths, and using subtle methods to influence others. Both involve deliberately manipulating the truth, making the boundary between them blurry. However, while lies are a type of deception, not all deceptive acts involve outright lies.

## 8 CONCLUSION

Our study, pivoting on the structural and linguistic nuances of Reddit dialogues, establishes a data-driven approach to discern critical sub-components like hate speech, trolling, and deception. We meticulously classify individual nodes 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. As a part of future study, we aim to expand our analysis across diverse social media platforms and further refine the computational models to encapsulate the evolving dynamics of online interactions. We believe that our findings could lay the groundwork for enhanced moderation tools and a deeper comprehension of discourse on digital platforms.

## REFERENCES

- [1] Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. *arXiv preprint arXiv:2005.04246* (2020).
- [2] Meredith D. Clark. 2020. DRAG THEM: A brief etymology of so-called “cancel culture”. *Communication and the Public* 5, 3-4 (2020), 88–92.
- [3] 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.
- [4] 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.
- [5] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [6] 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).
- [7] Luis Gerardo Mojica. 2016. Modeling trolling in social media conversations. *arXiv preprint arXiv:1612.05310* (2016).
- [8] Thomas S Mueller. 2021. Blame, then shame? Psychological predictors in cancel culture behavior. *The Social Science Journal* (2021), 1–14.
- [9] 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]
- [10] Diana Rieger, Anna Sophie Kumpel, 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.
- [11] Anna Vartapetian and Lee Gillam. 2014. Deception detection: dependable or defective? *Social Network Analysis and Mining* 4 (2014), 1–14.
- [12] 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.