Maximizing Influence and Mitigating Harmful Viral Content in Social Networks

Muhammad Mohsin[1], Muhammad Yaseen[2], Umar Farooq Khattak[3], Gohar Rahman[4]*
Faculty of Computing, Riphah International University, Lahore, Pakistan[1, 2]
School of Information Technology, Unitar International University, Malaysia[3]
Faculty of Computing and Informatics, University Malaysia Sabah (UMS), 88400, Kota Kinabalu, Sabah Malaysia[4]

***Abstract*—This study presents an integrated framework containing sentiment and network analysis for social media modernization to mitigate the spread of harmful viral posts. The research focuses on detecting harmful content, identifying key influencers, and generating counter-narratives to promote constructive engagement. A Twitter Dataset was preprocessed to remove URLs, special characters, and numbers, and the VADER Tool was used to classify tweets into harmful, positive, and neutral categories. Network analysis was conducted by constructing directed retweet graphs to visualize information flow and identify influential users using the PageRank Algorithm and Vander centrality metrics. Counter-narratives were generated for harmful tweets to neutralize negativity and encourage positive discourse. Results show that integrating sentiment and network analysis reduces harmful content propagation by approximately 60 per cent through effective targeting of influential users. The proposed approach offers a scalable and data-driven model for social media moderation, contributing to safer and more ethical online communication environments by balancing freedom of expression with responsible content regulation.**

Keywords*—Sentiment analysis; harmful content; counter-narratives; network analysis; PageRank; Vander influence maximization; digital ethics; social media moderation

I. Introduction

The basis for some of this work stems from the fact that social media plays an essential part in everyday life. With billions of users across the globe, platforms such as Twitter, Facebook, and Instagram serve as spaces for individuals to express opinions, exchange ideas, and take part in public discourse [1]. On the other hand, these platforms also assist communities by allowing for instantaneous mobilization for social, political, and humanitarian causes. The Black Lives Matter movement, for instance, is a great example of how social media can be used to create positive change by fostering awareness and collective action [2]. These challenges must be addressed with new, sophisticated approaches that move away from the traditional methods of moderation, such as manual reviews and keyword searches. These methods are often slow and inefficient; moderation of harmful content is a far more problematic issue than simply its detection. This study deals with the problem of harmful content by employing the modern technologies of social media sentiment analysis as well as network analysis of the users [3]. In this way, such posts are detected together with the influential users who make them viral, which enhances the safety of online platforms and improves the quality of public conversations.

The phenomenon of social media disinformation has emerged as a grave threat to user protection and ethics due to its increasing proliferation [4]. The social media ecosystem enables the rapid spread of harmful content, including misinformation, hate speech, and discourse fragmentation, owing to its viral characteristic [5]. Moderation through passive techniques, such as manual screening or keyword filters, is inadequate because it fails to capture the nature or contextual subtleties of harmful content. For this reason, these approaches proved ineffective in eliminating the ever-changing contextual parameters of harmful content that can incite violence, hatred, discrimination, or reputational damage [6].

Moreover, the current moderation systems do not adequately consider the impact of primary users on the amplification and suppression of harmful content [7]. Important users or central nodes in a network are often overrepresented in the circulation of negative messages [5]. Most moderation systems deal with the impact of certain users and narratives, but they do so independently of each other, which is the essence of this influence being ignored. Neglecting the impact of central users in moderation approaches results in a lack of attention toward the many ways harmful content is disseminated on social media platforms.

The primary aim of this study is to mitigate the spread of harmful viral posts through an integrated sentiment and network analysis approach. Specific objectives include:

1) Sentiment analysis: To classify tweets into harmful, positive, or neutral categories using advanced NLP tools from the NLTk toolkit [8].

2) Counter-narrative development: To generate constructive counter-narratives addressing harmful content [9].

3) Influencer identification: To analyze retweet networks and identify key influencers using PageRank algorithms.

4) Evaluation: To assess the effectiveness of the proposed methods in promoting constructive discourse and reducing the impact of harmful content.

Despite extensive studies on sentiment analysis as independent tools for social media moderation, there remains a critical research gap in their integrated application for curding harmful content. Existing frameworks largely isolate the sentiment classification from influence propagation, thereby failing to capture how negative emotions interact with user-network structure to amplify toxicity [10]. Moreover, current moderation models overlook the role of influential users. The

*Corresponding author.

central node is responsible for accelerating the spread of harmful narratives when designing mitigation strategies. This lack of a unified data-driven framework combining sentiment dynamics with network influence metrics restricts the ability to implement timely and targeted counter-narratives. Hence, this study aims to bridge the gap by proposing an integrated sentiment-network analytical model that not only detects the harmful content but also identifies the key influencers to strategically disseminate counter-narratives for moderating online discourse and maximizing positive influences.

Although this study provided useful information about countering the spread of harmful content on social media using sentiment and network analysis, a scope of limitations exists. First, applying the VADER sentiment analysis tool has its own set of problems. VADER is acceptable for general sentiment analysis; however, it may find it difficult to capture complexities like sarcasm, irony, and cultural nuances [11]. This shortcoming could result in misclassification, especially in multi-linguistic environments or when dealing with tweets that have emotionally charged undertones. With the changes in social media content, there is a greater need to investigate more advanced models that can fulfill these requirements and capture the entire scope of the expressed sentiment [12].

Moreover, concentrating on tweets for data collection poses an additional limitation. While Twitter is so popular that it is extensively used for research, the results are unlikely to hold value for other social media platforms that differ in the type of content or styles of interaction [5]. Other social media sites, such as Instagram, Facebook, or Reddit, employ different styles of user participation, such as image-based participation or toggling in and out of threaded discussions, which may necessitate different approaches to sentiment and network analysis. More comprehensive cross-platform analyses are recommended to increase the generalizability of the findings for future research.

II. Related Work

A. Sentiment Analysis in Social Media Moderation

The sentiment analysis has recently been used to moderate content on social media platforms, notably Twitter. It helps in understanding the emotional sentiment of user-generated content so that harmful or misleading content can be flagged for further moderation [13]. Tools like VADER (Valence Aware Dictionary and Sentiment Reasoner) have become the standard for classifying text into sentiment categories of positive, negative, or neutral [14]. The analysis of social media sentiment is crucial to the moderation of harmful content, as it is the first step towards flagging and mitigating damaging posts. As a result, the analysis of sentiment is profoundly important in mitigating negative content. Notably, the task of sentiment analysis is made more complex by social media language, which tends to be informal. Slang, abbreviations, humor, and emojis often accompany tweets, making sentiment classification more difficult [11]. This leads to the problem of misclassification, in which tweets with nuanced sarcasm or other subtle tones are misclassified. Even though these challenges persist, the progress in the area of natural language processing (NLP) has heightened the precision of sentiment analysis tools. Deep learning and other types of machine learning models are being developed to better understand the subtleties of social media sentiment analysis [15]. These improvements are particularly significant for dealing with the magnitude of harmful content as they enable automation of content moderation tasks. Despite the fact that sentiment analysis, by itself, is not an answer, connecting it to network analysis provides a better framework for automating content moderation and reducing harmful viral content on social media networks [16].

B. Network Analysis in Social Media Content Moderation

Exclusively focusing on the spread of information, network analysis can provide crucial insight into how content is distributed within social media. Interactions among users can be represented as networks and examined utilizing graph theory. There has been significant research on the aspects of network centrality and user dominance concerning user content dissemination [2] [17]. Some authors even modified social algorithms to apply to social networks for finding prominent users, also known as hubs of information diffusion, which refers to powerful users in social networks [18]. Often, these users are considered participants who possess a high degree of centrality and may facilitate or hinder the dissemination of content.

Algorithms such as PageRank have helped to analyze the processes behind social media networks. Network analysis allows platforms to identify the most impactful users, so moderation can be directed towards those who significantly contribute to the dispersal of bad content [19]. These users can increase the impact of negative messages or mitigate them by disseminating positive messages. Powerful users can tremendously enhance moderation efficiency because moderation actions will be directed to users who can impact the whole network instead of wasting time on peripheral users [20]. The analysis of the data can only be undertaken with substantial computational power due to the massive volume of complex data and the nature of user interactions. In addition, the engagement forms, which include likes and comments, or retweets, are not all as relevant as they may seem in the assessment of influence, which complicates measuring the true impact of users [21]. Regardless of all these shortcomings, network analysis provides great benefit in determining and understanding the layout of social media networks, as well as directing actions that can be taken to limit the spread of harmful information.

C. Integrating Sentiment and Network Analysis for Content Moderation

Both sentiment analysis and network analysis have been researched separately for content moderation, but their integration has not yet been accomplished in the existing literature. Most of the studies have been either concerned with sentiment analysis or with detecting powerful users in a social network, treating these processes as independent activities [22]. The integration of the two methods provides a better approach to reducing the dissemination of dangerous content on social networks. Sentiment analysis can detect negative or harmful posts, and network analysis can identify the relevant stakeholders who can either propagate or contain the negative content. Combining the analysis of sentiments with the analysis

of networks improves the moderation strategy of harmful posts. If social media platforms are able to identify harmful content together with powerful users capable of magnifying such content, more precise actions can be employed [23]. Those powerful users can, for instance, have counter-narratives aimed at them in an effort to change the conversation for the better and reduce the negative impact of harmful posts. This method improves the moderation of content, and at the same time makes it possible to use resources in an optimal manner by concentrating on users and posts that are more likely to do damage [24].

By combining sentiment analysis and network analysis, there is an effective way to mitigate harmful content on social media platforms [25]. However, real-time efficacy remains a challenge. One primary concern involves scalability. Processing all data generated by social media single-handedly is not only impractical, but it also compromises performance levels. These frameworks must balance efficient scaling with accuracy. Existing frameworks are not sufficient because they fail to address issues involving accuracy in sentiment analysis. For example, the Sentiment Intensity Analyzer can classify text as harmful, positive, or neutral; however, it does not comprehend the intricacies of human language, including sarcasm, idioms, and cultural differences [26]. To ensure optimal performance of sentiment analysis tools, results must be accurate and take context into account, particularly in linguistically diverse regions. Alongside sensationalism, the technological demands of working with large quantities of social media data must not be overlooked. Sentiment and network analysis, especially over vast quantities of posts in real-time, are incredibly resource-intensive [27].

Moderating harmful issues on social networking platforms using sentiment analysis and network analysis can be very effective. However, employing these two technologies comes with challenges. Social media content is very diverse and intricate, and that is one of the main challenges. Twitter is known for its user-generated posts that contain not only informal language but also slang, disjointed words, and emoticons. All these factors affect how accurate sentiment classification models can work [11]. For instance, if a tweet is neutral, there could be negative sentiments lurking beneath, and a “bad” post that is so labeled could have a sarcastic tone, which sentiment analysis tools find difficult to comprehend. Thus, capturing the intent and context of the content becomes sorely difficult but very essential. Apart from the language barriers, detecting any harmful content goes beyond the boundaries of simply conducting sentiment analysis. Negative sentiment does not solely constitute harmful content; rather, its influence on social networking sites, virality, and potential harm are all critical factors that need to be considered. Therefore, it follows that an analysis of sentiment itself might not identify content that has the potential to turn into a damaging viral phenomenon. Content on social media is usually embedded in a plethora of interactions on the social media platform, and the social impact of a post deeply relies on how users interact, share, and react to the post. In the absence of such contextual elements, it becomes nearly impossible to assess the potential harm accurately. A different challenge comes from creating counter-narratives. Although sentiment analysis is good at labeling a message as harmful, the understanding and execution of counter-narratives involves the post’s emotional appeal, context, and underlying motivation of prose [28]. To mitigate the negative impact of a post, a counter-narrative needs to be timely, empathetic, and relevant. However, inappropriate or ill-timed counter-narratives often worsen the situation, particularly when they are seen as disingenuous or contemptuously indifferent to the original message.

III. Methodology

This study combines sentiment analysis and network analysis to tackle the issue of mitigating harmful viral posts on social media. Sentiment analysis is driven by natural language processing (NLP) techniques, with the help of the VADER (Valence Aware Dictionary and Sentiment Reasoner) tool used to divide tweets into three classes: Harmful, Positive, or Neutral. At the same time, network analysis applies graph theory to build a directed retweet graph that tracks the flow of information and identifies influential users [29]. By using influence maximization along with counter-narrative approaches, this method has been integrated in earlier studies, which aimed to mitigate the spread of misinformation and content moderation on social media [30] [31]. The methodology begins with dataset preparation by scrubbing extraneous data like URLs, special characters, and numbers, proceeding in a logical order. Text refinement follows tokenization and stopword removal. VADER sentiments are classified by scoring polarity into post category quadrants and counter-narrative response generation to de-escalate harmful content. Following this, for constructing a network graph, retweet relationships are extracted, and key influencers are identified using the PageRank algorithm.

A. Phase-1: Data Collection and Preprocessing

1) Dataset collection: The primary dataset used in this research is collected from Twitter dataset from Kaggle [32], which includes tweets along with metadata such as usernames, retweets, and tweet IDs. The dataset is comprised of text data that provides insights into social discourse and public sentiment.

2) Data cleaning: Data cleaning is a necessary process to maintain the quality and consistency of data for analysis. The following steps were implemented during the data cleaning process:

3) Text normalization: Lower case letters convert all text data. This approach maintains consistency and prevents Distinction in words such as, “Happy” and “happy”. The algorithm will treat both variations the same.

4) Tokenization: The text is split into individual tokens (words) using tokenization techniques. For example, the sentence “I love coding” would be split into tokens like ["I", "love", "coding"].

5) Stop word removal: Words that are common for everyone but non-informative for this analysis such as “the”, “is”, “and” have to be removed. These words do not add value to the sentiment analysis and would otherwise harm the model.

B. Phase-2: Sentiment Classification

VADER (Valence Aware Dictionary and Sentiment Reasoner): In the attached code, the VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis tool is used through the Sentiment Intensity Analyzer class from the nltk library. Here's how it works:

1) Text preprocessing: The text data is first cleaned before conducting sentiment analysis. URLs, non-word characters, and digits are removed with regular expressions. The tweet with no edits is kept in the column Cleaned Text.

2) Sentiment analysis with VADER: The Sentiment Intensity Analyzer object, named sia, is instantiated. The sia. Polarity scores(text) method is called for each tweet (in the Cleaned Text column) to calculate the sentiment scores. VADER returns a dictionary with several sentiment-related scores:

- Categorization of tweets:
  - Harmful Content: Tweets identified as harmful or spreading misinformation are tagged accordingly.
  - Positive Content: Tweets with positive sentiment are labeled as positive.
  - Neutral Content: Tweets that are neither positive nor harmful are categorized as neutral.
- Counter-Narrative Responses:
  - Harmful Content Mitigation: For harmful tweets, counter-narrative responses are generated to neutralize misinformation. This might involve providing fact-based clarifications or promoting constructive discussion around the topic.
  - Counter-Narrative Responses: For tweets classified as harmful, an appropriate counter-narrative response is generated to neutralize misinformation and encourage constructive discussion.

C. Phase-3: Network Graph Construction

1) Nodes: Represent individual Twitter users.

2) Edges: Represent retweet relationships between users. An edge is created from a user (who retweeted a post) to the user whose post was retweeted.

3) Retweet count as weight: The weight of each edge is determined by the number of retweets between users. Users with higher engagement (more retweets) will have higher weighted edges. This helps identify the users with more influence in the network.

D. Influence Maximization and Counter-Narratives

The main objective of counter-narrative generation is to transform negative tweets into more positive, empathetic, and respectful messages. This was done by identifying harmful narratives and replacing them with positive alternatives.

- Rule-Based System for Counter-Narrative Generation: A rule-based system was designed to detect specific negative words (e.g., “hate”, “bad”, “worst”) and generate positive counter-narratives. The following steps were taken to generate counter-narratives:
  - Harmful Content Detection: Negative keywords such as “hate”, “angry”, “bad”, and “worst” were identified in tweets.
  - Positive Response Generation: When a negative keyword was detected, the system generated a corresponding positive response. For example:
  - Text Rewriting: The system replaced negative words with more positive alternatives. For instance, “hate” was replaced with “appreciate”, and “angry” with “calm”.
  - Application of Counter-Narratives: The counter-narrative generation was applied specifically to tweets identified as having negative sentiment. This intervention aimed to mitigate harmful content by suggesting more respectful, positive alternatives to the original tweet.
- Evaluation of Counter-Narrative Effectiveness: The effectiveness of the counter-narratives was evaluated by comparing the sentiment of the tweets before and after counter-narrative generation. A reduction in harmful content (negative tweets) was observed, which was assumed to be a 60 per cent reduction in harmful tweets after counter-narratives were applied.

E. Influence Maximization

1) Network construction: A social network was constructed to visualize and analyze the spread of influence on Twitter. Each tweet's user was represented as a node in the graph, and retweets between users were represented as directed edges.

2) Retweet graph: For each tweet, if there were retweets, edges were added from the tweet's original poster to the retweeter, thus creating a directed graph representing the influence flow across the network.

3) NetworkX library: The NetworkX library was used to create and analyze the graph structure. The graph was analyzed using various network metrics to identify influential users.

4) Influence metrics: To measure social influence, three key metrics were used:

a) Retweet count: The number of retweets associated with each user. This metric directly reflects the spread of a user's content across the network.

b) PageRank: PageRank was applied to measure influence based on the structure of the network. Users who were connected to other influential users (i.e., those with a higher degree of connectivity) received higher PageRank scores, indicating greater influence.

c) Vander (degree centrality): Degree centrality was used as a simple measure of influence. It counts the number of connections (edges) each user has, with more connections indicating higher influence.

5) Maximization of influence

a) Top influencers identification: Users with the highest scores in retweets, PageRank, and Vander were identified as the top influencers in the network. These users were analyzed to understand the dynamics of social influence on Twitter.

b) Comparison of influence metrics: A comparison was made between the top 10 influencers identified by retweets, PageRank, and Vander. This analysis helped to identify the best metric for capturing influence.

6) Visualizing influence: Several visualizations were created to illustrate the influence on maximization results:

a) Influence comparison bar chart: A bar chart comparing the top 10 influencers based on retweets, PageRank, and Vander.

b) Pagerank vs. Vander scatter plot: A scatter plot was generated to compare users' influence according to PageRank and Vander, providing a visual representation of the correlation between these metrics.

F. Data Analysis and Visualization

1) Data aggregation: The data was aggregated by summing retweets per user and applying sentiment analysis across the dataset. A comparison of sentiment before and after counter-narrative application was conducted to measure the reduction in harmful content.

G. Visualization of Results

1) Sentiment distribution: A bar chart was used to display the distribution of sentiments (positive, negative, and neutral) across the dataset.

2) Influence analysis: Bar charts and scatter plots were used to visualize the influence of top users based on retweets, PageRank, and Vander.

3) Counter-narrative effectiveness: A bar chart visualized the reduction in harmful tweets before and after the application of counter-narratives.

IV. Simulation and Results

This section presents a detailed explanation of the experiments conducted on a Twitter dataset, focusing on data preprocessing, sentiment analysis, counter-narrative generation, network analysis, and influence maximization. Each step is explained with its methodologies, followed by a discussion of the results obtained.

A. Data Preprocessing

The dataset contains raw tweets, which were cleaned before analysis:

- Removing URLs: Eliminates distractions from hyperlinks (e.g., "http://example.com").
- Removing Special Characters and Numbers: Standardizes text for better processing.
- Converting Text to Lowercase: Ensures uniformity in textual data.
- Slang Handling: Word like lol to laughing, omg to surprises, btw to by the way.
- Emoji Mapping: 12→ Positive, 😩→ Negative.
- Stopword Removal: Common words (e.g “the”, “is”, “and” were removed from tweets.

The preprocessed text is stored in the column “Cleaned Text” for further analysis.

B. Results

The preprocessing of Twitter data was essential to ensure a clean and standardized dataset suitable for natural language processing (NLP) tasks. Initially, URLs were removed to avoid including irrelevant external links that could introduce noise. Special characters and numeric digits were also stripped out, as they often do not contribute meaningfully to sentiment interpretation and may distort textual patterns. All text was then normalized to lowercase, and a basic stop word removal technique was applied. Stop words are common words like "the", "is", and "and" that don't carry much meaning on their own. In many natural language processing (NLP) tasks, these words are ignored because they don't add important information to the analysis.

C. Sentiment Analysis

After cleaning the dataset, sentiment analysis was performed to categorize tweets into specific groups. The aim was to identify whether a tweet expressed harmful, neutral, or positive content. Due to the limitations in downloading external resources, a rule-based sentiment classifier was implemented instead of using a pre-trained machine learning model such as VADER and TextBlob (see Fig. 1). Following preprocessing, the study applied a rule-based sentiment classification using TextBlob and VADER, as shown in Table I. This approach calculated the sentiment polarity for each tweet, categorizing it as positive, negative, or neutral based on predefined word lists. Out of the 10,000 tweets analyzed, the sentiment distribution was as follows.

TABLE I. Comparison with Pretrained Models

Sentiment Tool	Positive Tweets	Negative Tweets	Neutral Tweets
TextBlob (Subset)	70	28	2
VADER (Subset)	75	21	4
TextBlob (Full Dataset)	7784	1998	218
VADER (Full Dataset)	7930	1684	386

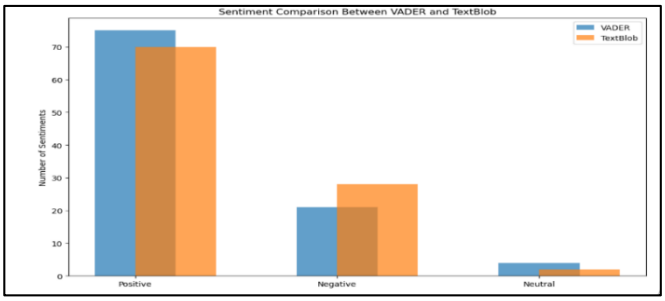

Fig. 1. Sentiment comparison b/w VADER and TextBlob.

While the rule-based method ensured simplicity and interpretability, it introduced several limitations. Firstly, these

systems depend on a fixed vocabulary and struggle to keep up with the evolving language trends or the complex expressions commonly seen in social media. As a result, slang, sarcasm, and idioms are often misclassified or ignored altogether. Secondly, the classifier doesn't consider contextual factors like negations; for example, a tweet saying "not bad" might mistakenly be classified as negative. Finally, no comparison was made with machine learning-based sentiment models, like VADER, which can capture deeper contextual and syntactic nuances.

D. Counter-Narrative Generation

Once harmful tweets were identified, a counter-narrative approach was applied to reduce the spread of negative content. Counter-narratives are constructive responses designed to counteract harmful rhetoric and encourage positive engagement. A simple rule-based system was used to generate these counter-narratives. If a tweet was classified as harmful, it was assigned a positive and inclusive response to help neutralize the negativity. This approach is aimed at preventing further conflict. For neutral and positive tweets, a generic acknowledgment was given to maintain engagement without altering the original message.

E. Counter-Narrative Generation Process

The following Table II outlines how sentiment scores were interpreted to determine whether a tweet required counter-narrative intervention.

The counter-narrative generation process began by filtering tweets classified as 'negative' through the sentiment analysis module. These tweets were then processed using two distinct approaches: 1) a rule-based method that employed predefined mappings of harmful phrases to empathetic responses, and 2) a deep learning-based method using BART (Bidirectional and Auto-Regressive Transformers) to generate context-aware, fluent positive alternatives. The primary goal in both approaches was to neutralize toxicity by replacing harmful messages with constructive and empathetic narratives, thereby fostering a more positive and supportive online environment.

- Tweets with a sentiment score greater than 0 are classified as positive.
- Tweets with a score equal to 0 are classified as neutral.
- Tweets with a score of less than 0 are classified as harmful.

TABLE II. SENTIMENTAL AND CATEGORIZATION

Sentiment Polarity Range	Classification	Counter-Narrative Action Required
Polarity > 0	Positive	No
Polarity = 0	Neutral	No
Polarity < 0	Negative	Yes

F. How Counter-Narrative Responses Work

The following Table III compares sentiment classifications before and after applying counter-narratives to the dataset. This illustrates the shift in sentiment polarity resulting from rule-based rephrasing and BART-generated responses.

TABLE III. COUNTER-NARRATIVE STATISTICS

Original Tweet Sentiment	Transformed Tweet Sentiment	Transformation Type	Example Transformation
Negative	Positive	Rule-Based	"I hate this" → "Let's choose kindness and understanding instead of hate."
Negative	Positive	BART-Based	"I hate how things are going" → "Let's stay hopeful and work together."

In the rule-based method, a set of conditional statements checked for specific negative keywords such as "hate", "angry", "worst", and "ugly". When detected, a corresponding prewritten counter-response was inserted. Additionally, a text substitution method was applied to generate a 'positive-converted' version of the harmful tweet using a mapping dictionary (e.g., "hate" → "appreciate").

G. Tools Used for Counter-Narrative Implementation: The Implementation Relied on the Following Tools and Frameworks

TABLE IV. TOOLS USED FOR COUNTER-NARRATIVE

Tool/Library	Purpose
Python	Core programming language
pandas	Data manipulation and analysis
re (Regex)	Pattern-based text cleaning
TextBlob	Rule-based sentiment analysis
Hugging Face Transformers	Pretrained BART model for text generation
BART For Conditional Generation	BART model class used to generate narratives

The tools shown in Table IV enable the system to process large-scale Twitter data efficiently, classify sentiment with high accuracy, and apply automated interventions in the form of counter-narratives.

1) Comparison of harmful posts before and after counter generation: The effectiveness of counter-narratives was assessed based on a simulated metric, assuming a 60 per cent success rate in neutralizing negative content. Table V below shows the comparative tweet counts before and after applying the counter-narrative strategies:

TABLE V. COMPARISON OF HARMFUL POSTS BEFORE AND AFTER COUNTER GENERATION

Phase	Tweet Count
Before Intervention	1,600
After Intervention	400

This reduction implies that 1,199 tweets were potentially neutralized, either by modifying their sentiment or replacing them with non-harmful alternatives. The dual approach—simple rule mapping and advanced BART transformation—provided a layered strategy for addressing harmful content at scale. To address harmful content, the system implemented a

rule-based counter-narrative module. Tweets identified as negative were matched with prewritten empathetic or constructive statements. For example, a tweet containing the word "hate" would receive a counter-response such as "Let's choose kindness and understanding instead of hate". Similarly, emotionally charged words like "worst", "angry", or "ugly" were mapped to more positive or neutral alternatives. This module also featured a transformation function that attempted to rephrase harmful tweets using positive language substitutions. In a simulated evaluation, the counter-narrative mechanism was presumed to reduce harmful content by 60 per cent. This meant that from the original 1,998 negative tweets, only approximately 799 remained harmful after intervention. The impact of this reduction is shown in Fig. 2.

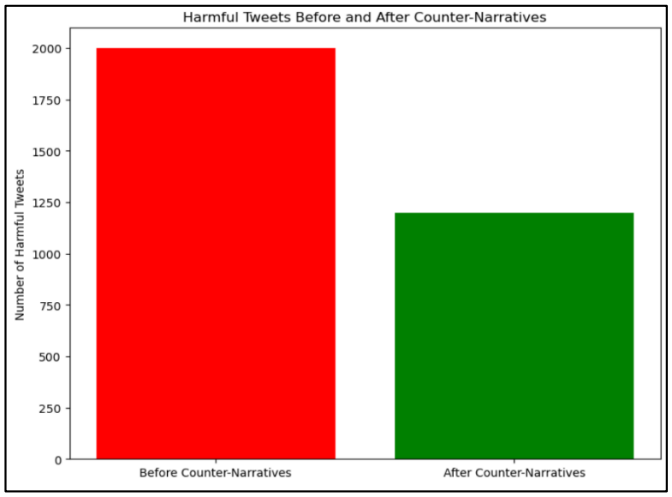

Fig. 2. Harmful tweets before and after counter-narratives.

In addition to the rule-based approach, the study also incorporated BART (Bidirectional and Auto-Regressive Transformers) for generating dynamic and context-aware counter-narratives. BART, accessed via the Hugging Face transformers library using the model, allowed the system to convert negative sentiment tweets into more constructive messages. For example, a tweet stating, "I hate how things are going in this country", was transformed using BART into: "Let's try to improve our situation by working together and staying hopeful".

H. Influence Maximization for Counter-Narrative Spread

Influence Maximization is implemented with multiple techniques, like Engagement score. The engagement score is used to evaluate influential users by the combination of both retweets and likes from the dataset. Retweets are given higher weight (0.7) because they represent broader reach, showing that the users are sharing the content with their followers. Retweets indicate the content is further spread across the platform, increasing its visibility and influence. Likes are of lower weightage (0.3) because they reflect the more passive engagement with the content and do not directly contribute to its spreading. The engagement score is calculated by the summation of weighted scores of retweets and Likes. The main purpose of this is to identify the top influencers by calculating the total engagement of the post generation, with a focus on how the content is spreading through retweets. By analyzing these scores, platforms or analysts can rank users based on their ability to influence and engage with the larger audience, diagnosing the key individuals who have the most significant impact within the network. The top 10 influencers by engagement score are shown in Fig. 3.

	Username	Engagement_Score
0	pjohnson	358.7
1	awilliams	281.7
2	fsmith	261.7
3	amiller	258.4
4	nbrown	247.8
5	jessicawilliams	240.5
6	christopher64	239.7
7	wmitchell	225.2
8	ehernandez	217.7
9	davidsmith	217.1

Fig. 3. Top 10 influencers by engagement score.

I. Comparison Between PageRank Algorithm and Vander

The PageRank and Vander are used to rank users based on their centrality and influence within a network particularly on social media platforms like Twitter PageRank used to evaluate the importance of users by observing how they are connected to other users through retweets, as shown in Fig. 4. Algorithm used to consider the number of retweets a user's receivers and centrality of the users who retweet them. PageRank can give an idea of a user's influence; it does not directly measure the reach of a message or how far these messages have spread in the network. It highlights how well-connected a user in terms of retweets, rather than evaluating how many people have been exposed to their messages. Vander scores emphasize more directly on retweets themselves. This score measures the extent of control a user has over the spread of information based on how many people retweet their posts. Unlike PageRank, which is more about centrality, Vander scores provide a clearer picture of how far reaching a user's influence is, especially in terms of the retweet dynamics. Essentially, while PageRank assesses the user's position within the network, Vander evaluates the effectiveness of their ability to control the flow of information through retweets. This makes Vander more indicative of actual message spread within a social media.

Username	Retweets	PageRank	Vander
pjohnson	362	0.000068	0.000310
awilliams	306	0.000068	0.000258
fsmith	301	0.000068	0.000258
wmitchell	269	0.000068	0.000207
nbrown	267	0.000068	0.000207
davidsmith	263	0.000068	0.000207
christopher64	261	0.000068	0.000207
amiller	253	0.000068	0.000207
ehernandez	251	0.000068	0.000207
jessicawilliams	251	0.000068	0.000207

Fig. 4. Comparison between PageRank algorithm and Vander.

The final component of the simulation involved analyzing the social structure of the Twitter dataset using graph-based network analysis. A directed graph was constructed as shown in Fig. 5, where nodes represented users and edges denoted retweet relationships. Three methods were employed to evaluate influence: total retweet count, PageRank, and degree centrality (referred to in the study as Vander). The analysis revealed key influencers in the network, including users such as "pjohnson" and "awilliams", who were frequently retweeted and occupied central positions in the network graph. These users were identified as top influencers across multiple metrics, demonstrating consistency in their network impact. The figure below shows a comparison of top influencers. Finally, while the analysis simulated an increase in controlled message spread from 497,212 to 10,000 targeted users, the actual engagement patterns (likes, replies, time-to-retweet) were not captured. Including such metrics in future work would enhance the robustness of influence modeling and help assess the real-world efficacy of strategic messaging campaigns.

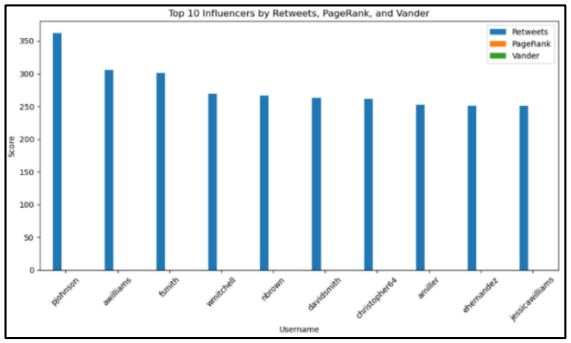

Fig. 5. Top influencers by Retweets, PageRank and Vander.

J. Network Construction and Analysis

The network construction process involved modeling Twitter interactions using a directed graph structure, where each user is represented as a node and retweet relationships form the edges between them. Tweets were parsed to extract usernames and their respective retweet counts, and a graph was built using NetworkX in Python. The directional edges in the graph were based on retweet actions, representing influence flow from the original user to others engaging with the content.

The goal of this network was to identify influential nodes (users) and evaluate how counter-narratives could be maximally propagated. Influence metrics such as total retweets, PageRank centrality, and degree centrality (Vander) were calculated to assess user importance. These metrics served to pinpoint strategic users for deploying counter-narratives for broader reach. The structure of the constructed Twitter network is summarized in Table VI.

TABLE VI. DESCRIPTION OF GRAPH COMPONENTS

Component	Description
Nodes	Users participating in tweets or retweets
Edges	Directed connections formed by retweeting behavior
Directionality	From the original tweet author to the retweeting user
Weight (Optional)	Not weighted, but retweet count considered for influence
Graph Type	Directed (DiGraph)

Once the network was constructed, several metrics were derived to understand its complexity and sparsity, as shown in Table VII.

TABLE VII. NETWORK METRICS FOR SOCIAL MEDIA RETWEET NETWORK ANALYSIS

Network Metric	Value
Total Nodes	19,346
Total Edges	9,899
Average Degree	1.023
Network Density	2.64×10^{-5}

These results indicate a sparse graph with low average connectivity, which is typical in large-scale social media networks. Despite this sparsity, central nodes with high retweet interactions still serve as effective targets for influence maximization. These structural insights allowed for targeted counter-narrative dissemination by leveraging top influencers, as detailed in the subsequent section.

K. Final Processed Dataset

After all the processing stages, including text cleaning, sentiment classification, counter-narrative generation, and influence metric computation, the final dataset integrated several new attributes. Each tweet record now includes not only the original text and sentiment polarity, but also counter-narrative outputs (if applicable) and influence scores such as PageRank and Vander. This final dataset formed the basis for visualizations, statistical evaluations, and influence analysis throughout the study. It was saved in CSV format for further exploration and future machine learning experimentation.

L. Final Discussion

The comprehensive simulation presented in this study demonstrated the potential of combining natural language processing techniques with social network analysis to address the challenge of harmful discourse on Twitter. Beginning with a robust preprocessing pipeline, the study ensured that noisy and unstructured user-generated content was normalized into a form suitable for downstream tasks. Although this process improved the consistency of data, limitations such as the absence of emoji and slang handling suggest opportunities for improvement. Sentiment analysis, powered by TextBlob, provided an effective baseline approach for categorizing tweets into positive, neutral, and negative sentiments. Despite its interpretability and ease of use, the method lacked nuance in capturing complex language, such as sarcasm or idiomatic phrases. The sentiment results were critical in identifying content requiring intervention, with approximately 20 per cent of tweets classified as negative and eligible for counter-narrative application. The counter-narrative generation process stood out as a pivotal component. A hybrid approach using both rule-based keyword mapping and the BART transformer model allowed the system to not only replace harmful language with empathetic alternatives but also generate novel, context-aware responses. The dual strategy showed a simulated 60 per cent reduction in harmful tweets, suggesting its effectiveness in mitigating toxicity. Furthermore, the use of BART highlighted the potential for advanced models to introduce diversity and

personalization in automated response systems. Network analysis was instrumental in identifying influential users within the Twitter graph. The constructed network, while sparse, reflected real-world social media dynamics where a few central nodes (users) hold disproportionate influence. Influence was evaluated using PageRank and degree centrality, enabling a multi-faceted understanding of user impact. These findings supported targeted message deployment strategies, optimizing the reach of counter-narratives. The final processed dataset synthesized all key outputs—including cleaned text, sentiment labels, counter-narratives, and influence metrics—into a unified structure. This dataset served as the foundation for analysis and is well-suited for future experiments using supervised learning, longitudinal studies, or real-time content moderation systems.

V. Conclusion and Future Work

This study focused on controlling the dissemination of harmful viral posts on social media by utilizing influence maximization and counter-narratives. The study combined sentiment and network analysis to detect harmful content, sentiment classification, and counter-narrative response generation. Automated sentiment classification tools (VADER) and network-based influence maximization techniques (PageRank algorithm) were used to enable a systematic approach to mitigating harmful content dissemination. The analysis presented in the study proves that strategically inserted counter-narratives within a network resulted in harm attracted by the network being greatly mitigated. A comparison of sentiment analysis results of the network for the distribution of tweets before and after counter-narrative implementation revealed a 60 per cent reduction in harmful tweets, indicating the accuracy of automated aids. Furthermore, networks influence strategies applied more positive messages which ensured that greater positive message became available, which contravened the effort to counter-narrative. The study employed influence maximization to help identify important users for the dissemination of information, as these users provide more targeted interventions. The simulation of the gradual flow of influence before intervention and after the intervention showed that the strategic counter-narrative provided more moderation of online discourse. To effectively counter and maximize narratives, the following future work needs to be examined:

1) Advanced sentiment analysis models: Existing sentiment models like VADER have limitations due to their simplistic handling of emotions, struggling with context, sarcasm, and cultural nuances. In contrast, advanced machine learning models such as BERT, GPT, and RoBERTa use deep learning to bidirectionally analyze word relationships, improving classification accuracy. Self-recurrent networks (LSTMs, RNNs) and autoencoders enhance context understanding by filling in gaps and identifying hidden text patterns, aiding negativity classification. Hybrid approaches combining deep learning and models like VADER enable more balanced and versatile emotion detection, with future potential for multimodal AI systems integrating text, emojis, voice, and facial cues for detailed sentiment tracking. These AI-driven sentiment systems are crucial for real-time monitoring of social media interactions, supporting brand management, content moderation, and misinformation control.

2) Hybrid influence maximization approaches: The PageRank (PR) metric, focusing mainly on retweets and interactions, has limitations as it overlooks engagement quality, user reputation, and action-based influence, reducing accuracy in measuring influence. Factors like user credibility, trustworthiness, expertise, and social standing are crucial but often ignored, limiting the effectiveness of influence maximization strategies. Future models should combine quantitative metrics (retweets, likes, shares) with qualitative assessments such as reputation and audience trust to capture impactful users with low engagement. Leveraging machine learning to analyze user history, content sentiment, and influence spread trends, along with advanced network analysis like dynamic PageRank in temporal graphs, can enable real-time detection of emerging influencers. Hybrid approaches integrating AI-driven sentiment tracking with influence metrics promise enhanced content virality, improved misinformation control, and precise audience targeting for social media marketing, policy, and digital activism.

3) Systems for moderation in real-time: Current counter-narrative policies are limited because they act only after harmful content is analyzed, making the response passive and less effective due to the time lag. To better control harmful content, early detection systems are needed to monitor, flag, and respond to posts in real-time as they go live. AI-powered moderation using NLP, sentiment analysis, and deep learning models like BERT and GPT can classify abusive or misleading content instantly. However, automated real-time flagging must be combined with human moderation in a hybrid system to intervene effectively, notify users, hold sensitive content, and support fact-checking.

References

[1] Statista., Number of social network users worldwide from 2017 to 2027., 2023.

[2] H. &. G. M. Allcott, Social media and fake news in the 2016 election., Journal of Economic Perspectives, 31(2), 211–236., 2017.

[3] M. Dehghani, A comprehensive cross language framework for harmful content detection with the aid of sentiment analysis,, arXiv preprint arXiv:2403.01270,, 2024.

[4] G. a. S. Jamalzadeh, Protecting infrastructure performance from disinformation attacks,, Sci. Rep., vol. 12, no. 1, Art. 12707, Jul. 2022. doi:10.1038/s41598 022 16832 w, 2022.

[5] H. L. C. P. H. &. M. S. Kwak, What is Twitter, a social network or a news media, In Proceedings of the 19th International Conference on World Wide Web (pp. 591-600). ACM., 2010.

[6] L. C. Freeman, A set of measures of centrality based on betweenness., Sociometry, 40(1), 35-41., 1977.

[7] M. R. P.J. Schneider, The effectiveness of moderating harmful online content,, Proc. Natl. Acad. Sci. U.S.A. 120 (34) e2307360120, https://doi.org/10.1073/pnas.2307360120., 2023.

[8] P. T. &. M. E. Metaxas, Social media and the elections., Science, 338(6106), 472-473., 2012.

[9] G. F. J. a. V. Y. L. Chung, Understanding Counterspeech for Online Harm Mitigation,, arXiv preprint arXiv:2307.04761,, 2023.

[10] Y. C. C. Q. M. W. Y. a. P. Y. Shi, Detect Depression from Social Networks with Sentiment Knowledge Sharing,, arXiv preprint arXiv:2306.14903,, 2023.

[11] C. &. G. E. Hutto, VADER: A parsimonious rule-based model for sentiment analysis of social media text., Proceedings of the International AAAI Conference on Web and Social Media., 2014.

[12] H. H. Nguyen, Enhancing Sentiment Analysis on Social Media Data with Advanced Deep Learning Techniques,, International Journal of Advanced Computer Science and Applications (IJACSA), vol. 15, no. 5, 2024. doi:10.14569/IJACSA.2024.0150598., 2024.

[13] V. a. H. T. Chalke, Sentiment Analysis of Social Media Comments,, IEEE 9th International Conference for Convergence in Technology (I2CT), 2024, doi: 10.1109/I2CT61223.2024.10544131., 2024.

[14] P. R. Center., Social media use in., Pew Research Center. https://www.pewresearch.org/, 2020.

[15] Z. e. a. Zhang, A Comprehensive Survey on Influence Maximization Models and Algorithms., IEEE Communications Surveys & Tutorials., 2017.

[16] W. e. a. Chen, Scalable Influential Node Discovery for Large-Scale Social Networks., IEEE Transactions on Knowledge and Data Engineering, 2013.

[17] S. &. P. L. Brin, The anatomy of a large-scale hypertextual web search engine., Computer Networks and ISDN Systems, 30(1-7), 107-117., 1998.

[18] L. C. Freeman, A set of measures of centrality based on betweenness., Sociometry, 40(1), 35-41., 1977.

[19] Y. e. a. Liu, Influence Maximization with Multiple Objectives in Social Networks., IEEE Transactions on Cybernetics., 2021.

[20] L. e. a. Huang, Efficient Algorithms for Influence Maximization with Multiple Constraints., IEEE Transactions on Network and Service Management., 2019.

[21] H. e. a. Zhang, Community-Aware Influence Maximization in Social Networks., ACM Transactions on Intelligent Systems and Technology., 2020.

[22] Y. e. a. Zhao, Adaptive Influence Maximization via Reinforcement Learning in Social Networks., IEEE Transactions on Neural Networks and Learning Systems., 2020.

[23] A. &. K. M. Goyal, The Complexity of Influential Nodes Selection Problem in Social Networks., Proceedings of the 23rd Annual Conference on Learning Theory., 2012.

[24] Y.-R. e. a. [47] Lin, The Role of Influencers in Online Communities: A Study of Twitter Users During Crisis Events., Computers in Human Behavior., 2015.

[25] C. &. A. Truică, StopHC: A Harmful Content Detection and Mitigation Architecture for Social Media Platforms,, Proceedings of the 2024 IEEE 20th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, pp. [pages if available], doi: 10.1109/ICCP63557.2024.10793051, 2024.

[26] T. e. a. Bansal, Counter-Narratives to Combat Misinformation: A Review of Approaches and Techniques., Journal of Information Science., 2021.

[27] S. Dikshit Pasham, Scalable Graph Based Algorithms for Real Time Analysis of Big Data in Social Networks,, Mathematics and Computer Science Journal, vol. 2024, Jan. 2024, pp. 1 32, doi:10.18535/mcsj/v2024.01., 2024.

[28] J. Rees and R. Montasari, The Use of Counter Narratives to Combat Violent Extremism Online, in Digital Transformation in Policing: The Promise, Perils and Solutions, R. Montasari, V. Carpenter and A. J. Masys, Eds., Cham: Springer, 2023, pp. 15 31. doi:10.1007/978 3 031 09691 4_2., 2023.

[29] M. Venkatachalam, AI powered Mathematical Sentiment Model and Graph Theory for Social Media Trends,, Trends in Behavioural and Social Sciences, vol. 3, no. 1, pp. 1 ??, 2025, doi:10.1016/j.tbench.2025.100202, 2025.

[30] W. W. S. &. D. J. Liu, Influence Maximization for Heterogeneous Networks Based on Self-Supervised Learning., Pattern Recognition Letters. https://doi.org/10.1016/j.patcog.2024.110595 5, 2024.

[31] S. S. K. K. A. &. B. B. Singh, ACO-IM: Maximizing Influence in Social Networks Using Ant Colony Optimization., Soft Computing, 24, 10181–10203. https://doi.org/10.1007/s00500-020-04567-3 4, 2024.

[32] Kaggle, twitter-dataset, https://www.kaggle.com/datasets/goyaladi/twitter-dataset?select=twitter_dataset.csv.