Harmful Meme Diffusion

by
Palani Vigneshwar

Under the supervision of
Dr. Md. Shad Akhtar
Dr. Tanmoy Chakraborty

Submitted in partial fulfillment of the requirements for the degree of Master of Technology, CSE-AI

Department of Computer Science Engineering
Indraprastha Institute of Information Technology - Delhi
May, 2023

Certificate

This is to certify that the thesis titled *"Harmful Meme Diffusion"* being submitted by **Palani Vigneshwar** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

May,2023 Dr Md. Shad Akhtar

Department of Computer Science Engineering
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

Acknowledgements

I am grateful to my advisor Dr.Md Shad Akhtar, and my co-advisor Dr Tanmoy Chakraborty for providing me with this opportunity of carrying out an independent research project and for allowing me to explore the topic of harmful meme diffusion. My advisors have provided support and guided me in the right direction throughout the duration of my thesis and academics.

I also thank my mentor Shivam Sharma who has provided me critical support during the challenging intervals of my research work.

This work would not have been possible without the constant moral and emotional support from my family and friends.

Abstract

With the widespread use of social media platforms, virality has become a fascinating subject for researchers. Content-based research emphasizes the role of content characteristics in predicting virality, such as the type of information presented, the emotional tone, and the format. On the other hand, Creator-based research looks at the characteristics of the person or group who created the content, such as their perceived credibility or popularity, as a factor in predicting virality. Both types of research can provide valuable insights into the factors that drive virality on social media. Individuals and organizations can create more effective social media campaigns and communication strategies by understanding the content and creator characteristics that promote sharing and diffusion. However, it is worth noting that many other factors contribute to virality, and the complex interplay between these factors can make it difficult to predict what will become popular online. One such factor is network-based features. We propose a topology of network features on top of content-creator features. We argue that adding the network features on top of the content-creator features will add conceptual richness and improve the predictive validity of future studies. We demonstrate this by running models, with and without the interactions, on a data set of nearly 100,000 posts from GAB, an American microblogging and social networking service known for its far-right user base. Our experiments show that our approach gives 92% F1 scores and ROC.

Contents

List of Figures

Chapter 1

Introduction

A meme is an element of cultural information, such as an image, video, or phrase, widely shared and spread rapidly across the internet. Memes often carry humorous or satirical content and are known for conveying ideas, emotions, or trends concisely and relatable. Individuals can create, modify, and share them, often reflecting current events, popular culture, or internet trends. Memes have become a significant part of online communication and have gained immense popularity and influence in contemporary digital culture.

A meme becomes viral when it rapidly spreads and becomes widely shared across various platforms and networks, gaining significant attention and engagement from many people within a relatively short period. The specific factors that contribute to the virality of a meme can vary, but some common elements include humor, relatability, novelty, and the ability to resonate with a broad audience. Memes that tap into current trends, events, or popular culture references often have a higher chance of going viral, as they can quickly capture the attention and interest of a wide range of individuals. Additionally, the shareability and ease of disseminating a meme through social media platforms and other online channels play a crucial role in its potential to become viral. One notable example of a viral meme is the “Doge” meme, which showcased a Shiba Inu dog accompanied by captions written in broken English and Comic Sans font. This meme originated in 2010 when a Japanese kindergarten teacher named Atsuko Sato shared pictures of her dog Kabosu on her blog. The meme gained widespread popularity and recognition because it inspired the creation of two cryptocurrencies named after it: the Shiba Inu coin and the Dogecoin. Another example of viral meme is The ”Overly Attached Girlfriend” meme, a popular internet meme featuring a still image or a series of images depicting a woman with an intense and possessive expression. The meme typically includes captions that humorously depict overly clingy or possessive behavior in romantic relationships. The image that sparked the meme originated from a YouTube video titled “JB Fanvideo”, uploaded by Laina Morris in 2012, in which she parodied Justin Bieber’s song “Boyfriend.” The video went viral, and a screenshot of Laina’s facial expression became widely used as the basis for the Overly Attached Girlfriend meme. The meme is often used to humorously exaggerate stereotypical behaviors associated with possessiveness or jealousy in relationships.

Over the last decade, numerous highly cited studies have been dedicated to studying online diffusion and virality (e.g., Berger and Milkman 2012[2], Susarla et al. 2012[14], Goel et al. 2015[5], Yu Han et al. 2019[6]). A study by Wu and Wang (2011)[17] discovered that positive online content about brands with higher source credibility leads to better attitudes toward the brand. Mills (2012)[10] proposed the SPIN framework, which considers content and creator characteristics when explaining virality but does not consider their interactions. Son et al. (2013)[13] variations in message characteristics observed based on the creator(poster) type, including firms, news media, and individuals. Hoang and Lim (2012)[7] and Yue et al. (2019)[6] developed a model that considers both tweet and user virality and found that it outperforms previous models in predicting retweet likelihood.

These studies have yielded valuable insights into the factors contributing to virality, primarily focusing on content and creator characteristics while giving limited attention to network features. Everyone belongs to one or more communities, which can be based on their location, workplace, preferences, etc. A strong community can facilitate organic reach and foster brand loyalty. However, it can also give rise to an echo chamber effect, wherein members are predominantly exposed to viewpoints that align with their own, reducing encounters with contrasting opinions or opportunities for critical thinking. Consequently, existing biases may be reinforced, impeding the expansion of perspectives among community members.

Our research note expands upon these studies by presenting a comprehensive nomological framework encompassing a broader range of interactions. Many memes are targeted explicitly toward particular individuals or groups, and their viral nature often relies on the audience’s familiarity with them. For instance, memes related to Donald Trump would likely garner significant viewership among American audiences due to their familiarity with him. In contrast, their reception would be less pronounced among African or South Asian audiences. To gain insights into which types of memes are likely to go viral within specific audience segments, it is crucial to identify the community affiliations of users. Therefore, understanding network characteristics is also essential in addition to content-creator features.

In the following sections we present the

1. Motivation

2. Related Work
3. Dataset
4. Proposed Methodology
5. Results
6. Future Scope

Chapter 2

Motivation

The topic of meme diffusion is important for several reasons:

1. Cultural impact: Memes play a significant role in shaping and reflecting popular culture. They can encapsulate current events, trends, and societal sentiments in a concise and easily shareable format. Understanding how memes diffuse allows us to comprehend their cultural impact on society better.

2. Communication and expression: Memes are a unique form of communication and expression in the digital age. They allow individuals to convey complex ideas, emotions, and humor in a relatable and shareable manner. By studying meme diffusion, we can gain insights into how messages and information spread online and how they resonate with different audiences.

3. Virality and influence: Memes often go viral, spreading rapidly across social media platforms and reaching a large audience. Examining the factors contributing to meme virality can provide valuable insights into how content becomes popular and influential digitally. This knowledge can be leveraged by marketers, content creators, and social media platforms to understand and harness the power of viral content.

4. Societal trends and behavior: Meme diffusion can offer insights into societal trends, attitudes, and behaviors. Analyzing the types of memes that gain traction and the communities that engage with them can provide a glimpse into different groups' preferences, beliefs, and cultural nuances. It can help identify emerging cultural phenomena and understand how ideas and ideologies spread among online communities.

5. Internet and media studies: Meme diffusion is a crucial area of study within the Internet and media studies field. It contributes to our understanding of online communication, social dynamics, and the ways in which information spreads in the digital realm. By exploring meme diffusion, researchers can gain a deeper understanding of the impact of digital media on society and the evolving nature of online interactions.

Chapter 3

Related Work

3.1 Content Features

Y. Han et al.[6] highlight the importance of studying the interactions between content and creator characteristics in predicting virality in social media. For content characteristics, the authors examined the length of the post, the presence of hashtags, mentions, and URLs, the use of emotional language, the sentiment expressed in the post, the category of the post, and the type of content (e.g., text, image, video). For creator characteristics, the authors looked at factors such as the number of followers, the number of followees, the age of the account, the frequency of posting, the race of the user posting through the report, and the topics of the previous posts by the user. R Rameez et al. focused on other creator characteristics, such as whether the user is verified or not and the average number of replies in the other posts of the particular content creator. The post text was also used for prediction by converting to BERT vectors.

3.2 Network Features

Weng et al.[16] investigate the relationship between virality prediction and community structure in social networks. The features used were early adopters, unaffected neighbors (set of users who can adopt the meme during the next step, infected communities (number of communities that the users who shared the meme belong to), adoption entropy based on infomap, topics that the members of the communities post on average, and proportion of the interaction regarding the meme belonging to the same community to the overall interactions.

3.3 Hate Features

Sharma et al.[12] define harmful memes as multimodal units consisting of an image and embedded text that has the potential to cause harm to an individual, an organization, a community, or society in general. Shraman et al.[11] introduce the MOMENTA framework, which utilizes a multimodal approach to analyze memes. The framework combines textual, visual, and contextual information to understand the memes' content and intent comprehensively. Text analysis involves natural language processing (NLP) methods to understand the textual content of memes, including the extraction of keywords, sentiment analysis, and identification of offensive or harmful language. Visual feature extraction focuses on analyzing the visual elements of memes, such as images or graphical elements, to identify potentially harmful or offensive content. This may involve image recognition, object detection, or visual sentiment analysis. The output of the process includes hate vectors and targets. Additionally, meme posts often contain text that can be potentially hateful. To quantify the toxicity of the text in these posts, we utilized Detoxify. Detoxify is a Python library developed by Unitary AI, available as an open-source resource. It offers pre-trained models specifically designed for detecting toxicity in the textual content. When provided with post text, Detoxify calculates the level of toxicity, obscenity, threats, insults, and identity attacks present in the text.

Chapter 4

Dataset

Gab.com is a social media platform that positions itself as a free speech platform, emphasizing a commitment to unrestricted speech and expression. It was created in 2016 by Andrew Torba as an alternative to mainstream social media platforms with perceived content moderation policies. Gab.com has gained attention for hosting a wide range of user-generated content, including controversial or extreme viewpoints that have been restricted or banned on other platforms.

Gab.com allows users to create profiles, follow other users, post messages (known as “gabs”), share images, and participate in discussions. It has been known to attract users with various political ideologies, including far-right and alt-right individuals. The platform has also been criticized for tolerating extremist, hateful, and harmful content, including white supremacist propaganda and hate speech. This has led to Gab.com being labeled as a platform that fosters hate speech and promotes extremist ideologies by some organizations and individuals.

The dataset was extracted using garc. Garc is a Python library and command line tool for collecting JSON data from Gab.com.

Type of Dataset	**Number of Rows**
Entire Gab Dataset	1615420
Original Posts with memes	99676
Number of users	161142

Table 4.1: Dataset Size.

The post-dataset consisted of 36 features. The list of important features is as follows:

- content: It has the post text, which was used for BERT vectors, sentiment analysis, and toxicity classification.

- in_reply_to_id: If the post was a reply, it consists of the original post details.
- media_attachments: It consisted of the URL of the meme.
- reblogs_count: It consisted of the total number of reblogs. It is used to get output for virality prediction.
- has_quote: If the post is a repost, it consists of the original post details.
- account: It consists of content creator details. The total number of posts by the user, whether the user is verified or not, and the account id is extracted from it..
- mentions: It consists of a list of users the content creator has mentioned(if any). Used to extract mentioned user count and in building network.
- tags: It consists of a list of hashtags the content creator has used(if any). Used to extract hashtags count.

The user dataset had 3 columns. These are:

- followers: Consists of the list of user ids that follow the mentioned user.
- following: Consists of the list of user ids the mentioned user follows.
- user_id: Mentioned user's id.

Chapter 5

Proposed Methodology

We divided the feature set into four categories. These are:

- Post Related Features
- BERT Features
- Image-Based Features
- Network Features

5.1 Post Features

Post-related features include the number of users mentioned in the post, the number of hashtags used in the post, the word length of the post text, the character length of the post text, the total number of posts made by the user, the number of followers of the users, four sentiment vectors obtained from Vader Sentiment on the post text, 24 topical affinity vectors based on the post, 24 topical affinity vectors based on the user's past posts, and six toxicity vectors based on detoxify applied on the post text. Detoxify is a toxic comment classification model that computes the post text into six vectors(refer 5.1). Topical Affinity Vectors determine the topic category for the post text. Zero-shot text classification was used on the post-text to obtain topical affinity vectors. The topics used in the affinity vectors are mentioned in Table Number 5.1. VADER(Valence Aware Dictionary for Sentiment Reasoning)[8] is an NLTK module that provides sentiment scores based on the words used. It is a rule-based sentiment analyzer in which the terms are generally labeled as per their semantic orientation as either positive or negative. VADER was applied on the post text to obtain four sentiment vectors(refer Table 5.1).

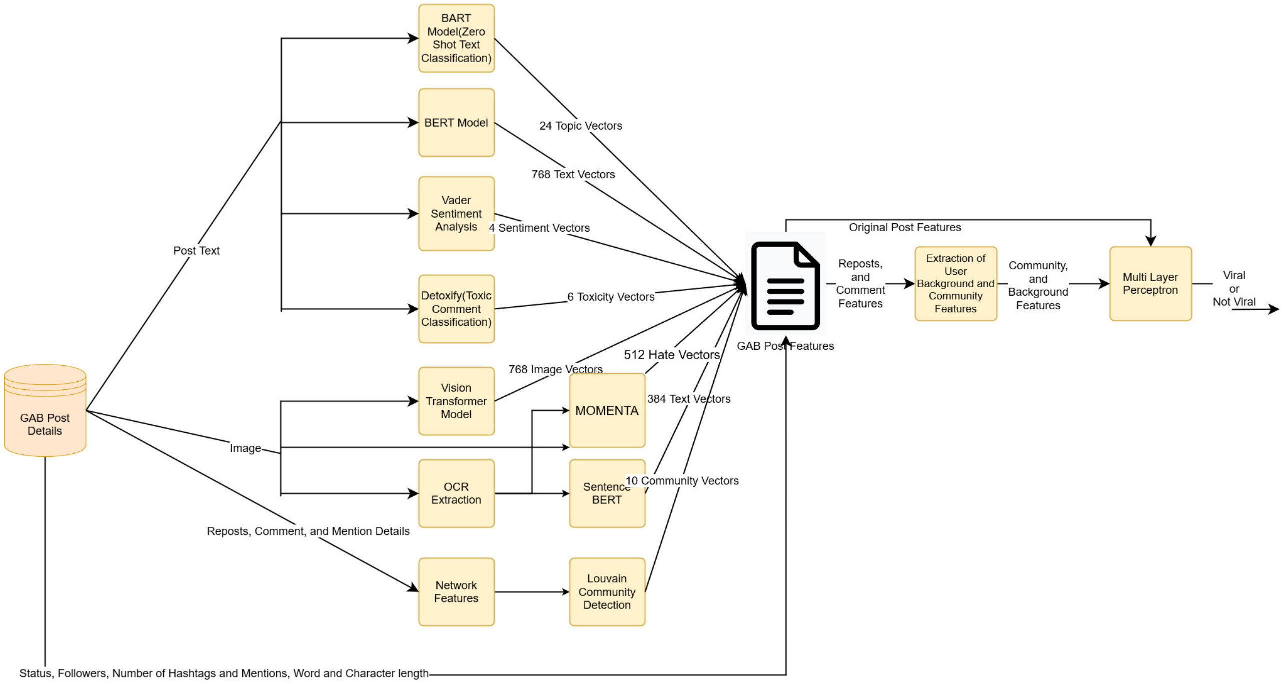

Figure 5.1: Methodology

5.2 BERT Features

Since post text size is large, the Longformer[1] model was used since transformer-based models cannot process large sequences due to their self-attention operation, which scales quadratically with the sequence length. The Longformer's attention mechanism scales linearly with sequence length, making processing thousands of tokens or longer documents easy. The Longformer model returns 768 vectors.

5.3 Image Features

The gifs were converted to images using Aspose, a Python library to convert gifs to images. The images were converted to 768 vectors using Vision Transformer[4] pre-trained on ImageNet-21k. The OCR was extracted using EasyOCR, a Python module for extracting text from images. For the images with no text, image captions were obtained from OFA. OFA[15] is a unified multimodal pre-trained model that unifies modalities (i.e., cross-modality, vision, language) and tasks (e.g., image generation, visual grounding, image captioning, image classification, text generation, etc.) to a simple sequence-to-sequence learning framework. The OCR and captions were then converted to 384 vectors using sentence bert. The image and the OCR text were then converted to 512 hate vectors using MOMENTA, a multimodal framework for detecting harmful memes.

5.4 Network Features

Each user is considered a node, and each mention, comment, and repost is considered an edge. This network is then passed through the Louvain community detection algorithm[3] to detect each user's community. Users who reacted to the post (i.e., commented or reposted) and belong to the same community as the poster is stored as SimilarInteractions. Betweenness centrality, Bridge nodes, density, and transitivity are stored based on communities. Using the all-pair shortest path method, we calculate the reachability of each node. We determine the early adopters and unaffected neighbors based on the users who have reacted to the post. The features are mentioned in Table 5.1.

Feature Type	**Feature Name**	**Total Number of Features**	**Sub-Feature Name**
BERT Features		768	
Post Features, Network Features	Toxify Vectors, Community Toxify Vectors	6	Toxicity, Severe Toxicity, Obscene, Threat, Insult, Identity Attack
Post Features, Network Features	Topic Affinity, User Background Topic Affinity, Community Topic Affinity	24	Asian, Black, Latinx, Middle Eastern, Native American, Pacific Islander, Race, Atheist, Buddhist, Christian, Hindu, Jewish, Mormon, Muslim, Religion, Immigrant, Transgender, Gender, Women, Bisexual, Gay, Lesbian, Sexuality, Disability
Post Features	Sentiment	4	Positive, Negative, Neutral, Compound
	Number of hashtags	1	
	Number of mentions	1	
	Word length	1	
	Character length	1	
	Number of posts	1	
	Number of followers	1	
Image Features	ViBERT Features	768	
	OCR Features	384	
	MOMENTA Features	512	
Network Features	Early Adopters	1	
	Unaffected Neighbours	1	
	Adopter Entropy	1	
	Reachability	1	
	Similar Interactions	1	
	Density	1	
	Transitivity	1	
	Infected Communities	1	
	Betweeness Centrality	1	
	Is Bridge Node	1	
	Community Size	1	

Table 5.1: Features

Chapter 6

Results

6.1 Baselines

6.1.1 Virality Prediction and Community Structure in Social Networks

The study[16] primarily emphasized network-related features instead of content or creator attributes. The network was constructed based on interactions such as retweets and mentions among the nodes. Community detection algorithms like Infomap and distributed Louvain were applied to analyze the network structure. Subsequently, various features were utilized following community detection, including:

- Early adopters: Number of users generating the earliest tweets.
- Unaffected neighbors: Followers of early adopters who can adopt the meme in the next step.
- Infected communities: Number of communities that contain the early adopters.
- Adoption entropy and usage entropy: The concept of entropy measurement pertains to the distribution of individuals who have embraced a particular meme among various communities.
- Fraction of intra-community user interactions: Pair-wise user interactions related to a specific meme are counted, and the proportion of these interactions occurring between individuals within the same community is calculated.

These features are then passed through a random forest algorithm that constructs 500 decision trees.

6.1.2 The Importance of Interactions Between Content Characteristics and Creator Characteristics for Studying Virality in Social Media

The paper[6] divides features into four categories:

- Content features
- Content message features
- Creator features
- Creator history features

Content features include:

- Length: Post Length.
- Hashtag: Use of hashtags. It could be zero if no hashtags are present or one if hashtags are present.
- Mention: Use of mentions to other users. It could be zero if no mentions are present or one if mentions are present.
- Link: Use of hyperlinks to external websites. It could be zero if no hyperlinks are present or one if hyperlinks are present.

Content message features include:

- Positive: Total number of positive words in the post text.
- Negative: Total number of negative words in the post text.
- Topic Affinity: Proportion of words about a topic in the post text. These topics are
  - Profane
  - Humor
  - Family
  - Pop
  - Politic
  - Sports
  - Relation

  - Animal
  - Emotion

Creator features include:

- TweetVolume: Total Number of posts.
- Followers: Number of followers.
- Following: Number of users the creator is following.

Creator history features include:

- Average post length: Mean length of all previous post text.
- Viral: Based on the virality criteria, the total number of previous tweets that went viral.
- Topic Affinity: Proportion of words about a topic in all the previous post text. The topics are similar to the topics of content message topic affinity.

These features are passed through the logistic regression model.

6.1.3 Dissecting the Meme Magic: Understanding Indicators of Virality in Image Memes

The paper[9] focuses on image characteristics to determine whether the post is viral or not. The features used here are:

- Number of Panels: The number of panels in the image. These are divided into single or multiple panels.

- Types of images: The images are divided into three types Photo, Screenshot or Illustration.

- Scale: Determines how the main subject is put in relation with the layout of the remaining elements of the image. These are divided into:

  - Close up
  - Medium shot
  - Long shot

- Type of Subject: The paper divided subject of the image into 4 types:

  - characters
  - scenes
  - creatures
  - objects

- Movement: The paper divided subject of the image into 3 types:
  - Physical movement: A hand in motion or people moving depicted in the image.
  - Emotional movement: Emotional expression on a face or body language depicted in the image.
  - Casual movement: The movement sequence is caused by one component (sender) to another (recipient).

- Attributes of the subject: For images whose subject is one or more characters, the paper considers whether the image's visual attraction lies with the character's facial expression or with their posture. These can be divided into:
  - Poster: Informative large scale image including both textual and graphic elements.
  - Sign: Informs or instructs the viewer through text, symbols, graph, or a combination of these.
  - Screenshot: Is a digital image that shows the contents of a electronic screen display.
  - Scene: A place where an event occurs
  - Unprocessed photo: Raw photo taken by a camera without being modified.

- Emotion: Emotions are divided positive, negative and neutral.

- Number of words

- Intended audience: The paper divides audience into two types:
  - Human Specific: Targetting a human.
  - Cultural Specific: Targetting a culture.

This is the passed through a random forest model.

S.No	Baseline	F1 Score	ROC Score
1	Virality Prediction and Community Structure in Social Networks	0.66	0.66
2	The Importance of Interactions Between Content Characteristics and Creator Characteristics for Studying Virality in Social Media	0.70	0.70
3	Dissecting the Meme Magic: Understanding Indicators of Virality in Image Memes	0.61	0.55
4	Proposed Model	0.92	0.92

Table 6.1: Baseline scores

6.2 Evaluation

6.2.1 Virality

Y. Han et al.[6] mentioned that for a post to be considered viral, the post's retweets should be greater than 50,000, and the retweets should be greater than or equal to two times the standard deviation of all the creator's previous tweet's retweets. However, since the total number of active users in GAB is less than that of Twitter (GAB has 100 thousand active users, and Twitter has 450 million active users), for a post in GAB to be considered viral, the reposts of the post should be greater than 100, and the reposts should be greater than or equal to two times the standard deviation of all the creator's previous post's reposts.

6.2.2 Experiment

The dataset is then oversampled using smote. The features mentioned in the table are passed through two models:

- Multi-Layer Perceptron with the maximum number of iterations as 1000 and early stopping.

- Logistic Regression with the maximum number of iterations as 10000.

Features	MLP		Logit	
	F1	ROC	F1	ROC
BERT Features	0.65	0.65	0.61	0.61
Post Features	0.81	0.81	0.63	0.63
Image Features	0.73	0.73	0.61	0.61
Network Features	0.70	0.69	0.59	0.59
Post Features + BERT Features	0.68	0.68	0.65	0.65
Network Features + Post Features	0.92	0.92	0.73	0.73
Network Features + Image Features	0.75	0.75	0.64	0.64
Image Features + BERT Features	0.72	0.72	0.65	0.65
Image Features + Post Features	0.75	0.75	0.67	0.67
Image Features + Post Features + BERT Features	0.73	0.73	0.67	0.67
Network Features + Image Features + Post Features	0.81	0.81	0.78	0.78
Network Features + Image Features + BERT Features	0.72	0.72	0.64	0.64
Network Features + Post Features + BERT Features	0.73	0.73	0.75	0.75
Network Features + Post Features + BERT Features + Image Features	0.75	0.75	0.70	0.70

Table 6.2: Feature Results

6.2.3 Observations

Based on our observation mentioned in Table , individually Post Features with Multi Layer Perceptron gives the best performance with 0.81 F1 Score and 0.81 ROC. This is followed by Image Features with Multi-Layer Perceptron with 0.73 as F1 score and 0.73 as ROC. Network Features with Multi-Layer perceptron gives 0.70 F1 score and 0.69 ROC. BERT Features perform poorly, having 0.65 as F1 Score and 0.65 as the ROC score with Multi-Layer Perceptron. The best model is Post Features and Network Features, with Multi-Layer Perceptron having 0.92 ad F1 Score and 0.92 as the ROC score. This surpasses the second-best model by 11% and surpasses the best baseline of Y. Han et al. by 22%.

6.2.4 Conclusion

Post text individually gives the best accuracy compared to other features. This can be attributed to features such as:

- is_verified: The verification status of a content creator on GAB, particularly among well-known right-wing personalities, plays a crucial role in determining the potential virality of their posts. Verified content creators, symbolized by the verification sign, often enjoy a larger following and higher engagement rates. As a result, their posts tend to attract considerably reblogs and shares, increasing the likelihood of them becoming viral within the GAB community. Verification indicates the credibility and influence of the content creator, amplifying the reach and impact of their posts.

- Number of posts: Frequent posting on GAB often indicates a user's high activity level on the platform. While this behavior is not limited to renowned individuals alone, it can serve as an additional factor suggesting the potential virality of a post. Regular and consistent posting patterns increase visibility and engagement, making it more likely for a post to gain traction and go viral among the GAB community.

- Followers: The number of followers a content creator possesses directly impacts the post's potential reach and virality. As the count of followers increases, more users gain access to the content shared by the creator. Therefore, the number of followers is a crucial indicator of a post's likelihood to become viral, as it reflects the broader audience that can potentially engage with and amplify the post.

- Toxicity vectors: Toxicity vectors significantly influence the potential virality of posts on GAB, an alt-right website known for its prevalent offensive and toxic content. The platform's nature and user base contribute to a heightened likelihood of posts with toxic elements gaining traction and spreading rapidly among the community.

- Number of hashtags: The usage of hashtags serves as a valuable indicator for predicting the virality of a post on GAB. This is due to GAB's feature that allows users to browse and explore posts associated with specific hashtags. Consequently, the more hashtags employed in a post, the higher the likelihood of it being exposed to more spaces within the platform. As a result, the post's visibility expands, increasing its reach and potential for virality. A higher number of hashtags correlates with a broader exposure and enhanced chances of a post gaining widespread attention and engagement.

Post Features and Network features, along with Multi-Layer perceptron, gives the best performance since network features have associated features in the form of community-based features. These features are:

- Community-Based Topic Affinity Features: If a community frequently posts on certain topics and the user post relates to one of the topics, the post will be more likely to spread across the community.
- Community-Based Toxicity Vectors: If a community frequently posts highly offensive posts. The user's post with similar toxicity is more likely to spread across the community.

Along with these associated features, there are other features that contribute to virality. These are:

- Is_bridge: Bridge users refer to individuals with connections to multiple distinct communities or groups within the network. These users act as connectors or intermediaries between different communities, facilitating the flow of information, ideas, and interactions across otherwise separate or disconnected groups. So if a user is a bridge user, his posts can be spread across multiple communities.
- Betweenness Centrality: Betweenness centrality is a measure used in network analysis to quantify the importance or influence of a node within a network. It calculates the extent to which a node lies on the shortest paths between other pairs of nodes in the network. Nodes with high betweenness centrality significantly influence the flow of information, as they act as critical bridges or connectors between different parts of the network.
- Community Size: The size of a network community in a social network can significantly influence the potential virality of a social media post. Generally, larger network communities provide a larger audience and more potential viewers for a post. When a post is shared within a sizable community, it is more likely to be seen, interacted with, and shared by a larger number of individuals. This increased exposure amplifies the chances of the post gaining traction, spreading rapidly, and ultimately becoming viral within the community.
- Reachability It represents the degree of separation between nodes and their proximity regarding network connections. Nodes with lower reachability, which can be reached with fewer hops, are more easily accessible within the network.
- Infected Communities: The reachability of a social media post refers to the total number of communities it has initially infected. The more communities a post infects, the wider its potential audience becomes, increasing the likelihood of the post going viral. By reaching a diverse range of communities, the post gains

exposure to a larger pool of users who may engage with it, share it further and contribute to its virality.

Chapter 7

Future Scope

7.1 Community Structure

The results demonstrate that network features, while not achieving the highest accuracy on their own, complement post features to produce the best outcome. However, the assumption underlying the computation of network features is that a user can only belong to a single distinct community. In reality, this assumption does not hold true, as individuals often participate in multiple communities, and these communities can overlap.

Taking this into account, it opens up a fresh perspective for examining the network. We can incorporate associated community-based features, such as community-based topic affinity and community-based toxicity, by considering them as weighted sums rather than simple averages of the post patterns across communities. This approach acknowledges the complexity of users' affiliations and allows for a more nuanced understanding of the network dynamics, potentially leading to improved performance and insights.

Bibliography

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[2] Jonah Berger and Katherine L Milkman. What makes online content viral? *Journal of marketing research*, 49(2):192–205, 2012.

[3] Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. Generalized louvain method for community detection in large networks. In *2011 11th international conference on intelligent systems design and applications*, pages 88–93. IEEE, 2011.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. The structural virality of online diffusion. *Management Science*, 62(1):180–196, 2016.

[6] Yue Han, Theodoros Lappas, and Gaurav Sabnis. The importance of interactions between content characteristics and creator characteristics for studying virality in social media. *Information Systems Research*, 31(2):576–588, 2020.

[7] Tuan-Anh Hoang and Ee-Peng Lim. Virality and susceptibility in information diffusions. In *Proceedings of the international AAAI conference on web and social media*, volume 6, pages 146–153, 2012.

[8] Clayton Hutto and Eric Gilbert. 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*, volume 8, pages 216–225, 2014.

[9] Chen Ling, Ihab AbuHilal, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Dissecting the meme magic: Understanding indicators of virality in image memes. *Proceedings of the ACM on human-computer interaction*, 5(CSCW1):1–24, 2021.

[10] Adam J Mills. Virality in social media: the spin framework. *Journal of public affairs*, 12(2):162–169, 2012.

[11] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*, 2021.

[12] Shivam Sharma, Firoj Alam, Md Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, Tanmoy Chakraborty, et al. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274*, 2022.

[13] Insoo Son, Dongwon Lee, and Youngkyu Kim. Understanding the effect of message content and user identity on information diffusion in online social networks. 2013.

[14] Anjana Susarla, Jeong-Ha Oh, and Yong Tan. Social networks and the diffusion of user-generated content: Evidence from youtube. *Information systems research*, 23(1):23–41, 2012.

[15] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[16] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3(1):1–6, 2013.

[17] Paul CS Wu and Yun-Chen Wang. The influences of electronic word-of-mouth message appeal and message source credibility on brand attitude. *Asia Pacific Journal of Marketing and Logistics*, 2011.