Project_Report_3.pdf

By Rahul Saxena

WORD COUNT	4970	TIME SUBMITTED	19-MAY-2020 09:55AM
		PAPER ID	58867461

Graph Based Approach for Analysis of Online Hate Speech

A dissertation report submitted for fulfilment of the requirements of the degree of B.Tech.

by

Uditi Arora (2016UCP1415)

Arnav Loonker (2016UCP1016)

Rishabh Kalakoti (2016UCP1098)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY

May 2020

Certificate

We,

UDITI ARORA (2016UCP1415)

ARNAV LOONKER (2016UCP1016)

RISHABH KALAKOTI (2016UCP1098)

Declare that this thesis titled, "Detection and Analysis of Hateful User on Twitter" and this report is of our own work. We confirm that:

- This project work was done wholly or mainly while in candidature for a B.Tech. degree in the department of computer science and engineering at MNIT.

- Where any piece of this postulation has recently been submitted for a degree or some other capability at MNIT or some other organization, this has been plainly expressed. Where we have counseled the distributed work of others, this is in every case plainly credited. Where we have cited from, the source is given. Except for such citations, this Dissertation is completely our own work.

- We have recognized every single primary source of help.

Signed: _____

Date: _____

Dr. Mahipal Jadeja

Assistant Professor

May 2020 Department of CSE

MNIT, Jaipur

Abstract

These days, hateful speech via web-based networking media have become a serious issue that does unpredictable damage to individuals, particularly youthful understudies and experts. To take care of this difficulty it is pivotal to effectively discover hateful users who convey disdainful addresses and can end up being unsafe for the online society. Many of the current ways to identify hateful speeches center around content posted in OSNs. They face inadequacies to gather and comment on hateful speeches because of the noiseiness and incompleteness of OSN text and the subjectivity of hateful speech. These restrictions are regularly supported with imperatives that distort the issue, for example, considering just tweets containing loathe related words. In this work we mostly address these issues by moving the concentration towards users. We use a Twitter dataset of 100, 386 users along a sample of 4, 972 manually annotated nodes. We have seen that hateful users are more compactly connected and are more central than other users. The subjectivity of their content is lesser than normal users but their engagement is comparatively higher. We also try to examine user activity in order to get insightful findings about hateful users. We use the sampled graph to evaluate network centrality measure, user activity and content. All our findings point towards the use of semi-supervised learning algorithms which support learning of not only user characteristics but also the properties of the neighbours. We also focus on exploiting the network connections of the graph. These are not connections related to followers or following but of retweeting. It can also be said that the retweet connections tell us more about the hateful behaviour rather than follower-followee relation. From observations and previous research it has been shown that twitter retweet graph is a **small-world model**. We used NetworkX to visualise the graph and preprocess data. We start from graph networks and implement and evaluate semi-supervised GraphSAGE. We also explore different versions of Graph Convolutional Networks (GCNs).

Acknowledgements

Foremost, we would like to thank our guide, Dr. Mahipal Jadeja, for his continuous guidance and support. He has held our hand throughout the process of this project and we couldn't have asked for anything more. We would also like to thank our HoD, Dr. Pilli Emmanuel Shubhakar, and the Computer Science department for giving us the opportunity of doing this project.

Uditi Arora	Arnav Loonker	Rishabh Kalakoti
(2016UCP1415)	(2016UCP1016)	(2016UCP1098)

Table of Contents

Certificate **2**

Abstract **3**

Acknowledgements **4**

Table of Contents **5**

List of Figures **7**

List of Tables **7**

Chapter 1 - Introduction **9**

Chapter 2 - Important Terms and Concepts **11**

2.1.1 Box Plots 11

2.1.2 KDE Plots 11

2.1.3 Network Centrality 11

2.1.4 Network Connectivity 11

2.2 Terms Related to Dataset 11

2.2.1 Hateful Users 11

2.2.2 Retweet Graph 11

Chapter 3 - Literature Survey **12**

3.1 Belief Propagation 12

3.2 GraphSAGE 12

3.3 Graph Attention 12

3.4 LoNGAE 12

3.5 GraphMix 13

Chapter 4 -Related Work **14**

Chapter 5 - Proposed Method **15**

Chapter 6 - Experimental Results **16**

6.1 Results of Experimental Data Analysis 16

6.1.2 Date of creation of accounts of hateful users 16

6.1.3. Behaviour of Hateful Users 17

6.1.4. Comparison of hateful users with spammers 18

6.1.5 Centrality of Hateful Users 18

6.1.6 Vocabulary used in hate speech 19

6.1.7 Sentiment of Hateful Users 20
6.1.8 Connectivity of Hateful Users 20
6.2 Models Used For Training 21
6.2.1 GraphSAGE 21
6.2.2 Graph Convolutional Network(GCN) 22
6.2.3 Hierarchical Graph Convolutional Networks(H-GCN) 23
6.2.4 Bayesian Graph Convolutional Neural Networks(BGCN) 24
Chapter 7 - Conclusion **28**
References **29**

List of Figures

6.1. Creation date of users …….. …….....17

6.2. Nature of users …….....17

6.3. Hateful users v/s spammers ……18

6.4. Centrality of Hateful Users ……19

6.5. Vocabulary of hateful users ……...
.19

6.6. Sentiment of hateful users ……...20

6.7. Plot of hateful users and how they are connected ……...20

6.8. Steps in GraphSAGE …… ……21

6.9. GraphSAGE loss v/s epochs …….....21

6.10. Basic GCN architecture ……22

6.11. GCN accuracy and loss v/s epochs ……...23

6.12. HGCN architecture ……...23

6.13 HGCN loss and accuracy vs epochs……...24

6.14. Graph edges in non-bayesian approach ……24

6.15. BCGN loss and accuracy vs epochs ……27

List of Tables

Table 1. Users in different groups ……16

Table 2. Model results ……..27

Chapter 1 - Introduction

The process of analysing hateful speech in Online Social Networks (OSNs) is complex. The trade-off between prohibiting such conduct from stages and not editing contradicting sentiments is a significant cultural issue. Recently, its impact can be seen on the conditions of the COVID-19 pandemic. Hateful speech has effects on the way people perceive major decisions taken by the government. It has a major hand in creating chaos during such serious times. These incidents have motivated people to work on understanding hateful content in Online Social Networks. The work also focuses on drawing insights on the nature of hateful speech. Although, in these Online Social Networks, meaning and context of hateful speeches is affected by recent events due to the presence of informal language, sarcasm, errors and special characters. For instance,

These problems, although observed, remain unaddressed. Take into account this tweet:

Time is up, you all are getting what should have been done centuries ago.

This tweet was reply to a tweet which talked about holocaust. The thing to be noticed is that this tweet doesn't feature any hate words but the writer's profile indicated that he was a white man who believed in superiority of white people and in violence. Moreover, our problem starts when we notice that there isn't a single hate related word and thus it's extremely hard to sample this kind of tweet. Thus, the solutions to these problems are still far fetched. Sampling of such tweets still remains a major issue while using NLP-like techniques to detect hateful speeches.

The information in posts, messages and tweets are not by any means the only signals we may use to examine hateful behaviour in Online Social Networks. More often than not, the signals of such hateful content are linked to the profile of a person and their profile represents the community they belong to, the values they hold, their beliefs and thus their likes and dislikes. A person's tweet says way less about him than his/her profile. So, we can conclude that detecting hateful users is more important than detecting hateful content/ tweets. While it has been believed that hateful tweets will lead us to hateful users, there is potential in the possibility that hateful users can lead us to hateful tweets which are disguised as normal ones. In practicality, this is expected to yield better results as this aims at eliminating the root of the problem. It is only natural that the users should be kept in check as the online world is becoming more and more dangerous by the day. More benefits of this approach include that we explore not only content but also user's activities and connections. It gives us a better perspective of the user's mindset and thus enhances our ability to detect similar users. We use Twitter's network structure and exploit its potential in trying to solve some major issues on this topic.

Breaking down hateful users as opposed to content is additionally appealing on the grounds that different measurements might be investigated, for example, the user's action and associations in the system. For instance, in Twitter, it is conceivable to see the quantity of tweets, supporters, and top choices a user has. It is likewise conceivable to separate impact joins among users who retweet one another, examining them in a bigger system of impacts. This permits us to utilize arrange based measurements, for example, betweenness centrality and furthermore to dissect the area of such clients. Recognizably, albeit a few examinations describe hateful behaviour in content, not many investigations center around the component of users in OSNs.

Our outcomes show that hateful users tweet more often, and most loved different tweets essentially

more than the normal users. They likewise are progressively negative as per vocabulary based supposition investigation and utilize more swear words. They also follow a greater number of individuals every day than typical ones, and use jargon identified with classifications, for example, loathe, outrage, disgrace, savagery and psychological oppression less as often as possible. Likewise, the median hateful user has higher system centrality as per our observations, repudiating the "lone wolf" conduct regularly connected with the training.We experimented with different algorithms based on node classification using graph networks. First we used the GraphSAGE semi-supervised classifier. We also used variations of GCNs, out of which BGCN gave the best results.

Chapter 2 - Important Terms and Concepts

2.1 Terms used in graphs plotted

2.1.1 Box Plots

A box plot is the type of graph which shows distribution of data in the form of a five number summary. It is a standardized way of representing data divided as: minimum, first quartile, median, third quartile and maximum. [1]

2.1.2 KDE Plots

KDE Plot described as **Kernel Density Estimate** is used for visualizing the Probability Density of a continuous variable.[2]

2.1.3 Network Centrality

Network centrality refers to the concept of graph theory which measures the centrality of a node in a network. It finds the most important and central vertices of a graph.

2.1.4 Network Connectivity

Network connectivity depicts the broad procedure of associating different pieces of a system to each other, for instance, using switches, switches and portals, and how that procedure works.[5]

2.2 Terms Related to Dataset

2.2.1 Hateful Users

Hateful users have been described according to the guidelines provided by Twitter . Here, hate speech refers to any content that attacks or threatens a person or promotes violence in any form against the community based on gender, gender orientation, religion, age, disability, nationality, sexual orientation, race, ethnicity, etc.

2.2.2 Retweet Graph

Retweet graph refers to directed which in which each node is a user and each edge from node u1 to node u2 means that user u1 retweeted user u2's tweet. [6]

Chapter 3 - Literature Survey

3.1 Belief Propagation

Considering real life scenarios, individuals via web-based networking media are firmly associated with individuals of comparative interests and foundations. Based on this assumption, the belief propagation algorithm was implemented. It is a baseline model which is structure based. It gave an accuracy of 0.8282. However, the labelled dataset which was used was highly unbalanced. 90% of the data were normal users. Due to this there is a need to incorporate node attributes as we cannot get much information about hateful users from such a small percentage of data.[7]

3.2 GraphSAGE

A GraphSAGE semi-supervised classifier was implemented using node classification with the help of Pytorch geometric. GraphSAGE is a framework that makes use of node features and neighbourhood information in the graph to generate node embeddings for data that had not been seen before. With this generated information in the form of embeddings, we conducted a classification for our dataset. The process also included preprocessing of data in the form of normalization and sampling. GraphSAGE-0 performed even worse than the baseline model of belief propagation with an accuracy of 0.8028. 25 neighbours of a labeled node were sampled for GraphSAGE. Using this variation took the accuracy up to 0.8489. Across all models, GraphSAGE with 25 neighbour sampling performs best among all models. [7]

3.3 Graph Attention

Graph Attention(GAT) classifier makes use of a mechanism based on biasing more relevant parts of information from the neighbours. The model trained here used 5 fold cross validation along with a Graph Attention classifier. It performs slightly worse than GraphSAGE as the accuracy drops to 0.8081. The exhibition diminished in light of the fact that the tests were directed utilizing one head, and it's presumably not incredible enough to learn numerous helpful highlights. [7]

3.4 LoNGAE

LoNGAE is used for graph link prediction. LoNGAE performs encoding and decoding transformations for link prediction. The autoencoder is learned with symmetrically shared parameters. In order to use the model for node classification, a feature set had to be added manually with a Softmax activation

layer. The input was both labeled data and sampled unlabeled data. The model was using labeled and sampled data. The results show that the accuracy was 0.8480. It was also observed that the performance goes down as soon as the number of unlabeled nodes go above 1000. This might be due to the fact that as unlabeled data increases, the algorithm becomes inefficient and doesn't record feature relationships. It also fails to give consistent performance and thus is definitely not reliable for use in larger networks. [7]

3.5 GraphMix

The GraphMix algorithm is based on node classification using GNNs for semi supervised learning.It prepares a diagram neural system and a completely associated organize by means of parameter sharing, addition based regularization and self-anticipated targets. It gave an accuracy of 0.884. It was observed with increasing time the performance dropped because the network became too big.[7]

Chapter 4 -Related Work

In previous works related to twitter, analysis was carried out on the follower graph. This graph connected users with each other on the basis of who follows whom. (Bild, 2015[8]) were the first ones to take a different approach and consider studying the retweet graph and deriving insights from it. They found out that Twitter retweet graph is a small world model and used decision tree classifier to do spammer detection on the dataset. (Ribeiro, 2017[9]) was the one who created the dataset which has been used in our project and is also available on Kaggle. He used matplotlib to characterise users on the basis of their sentiment and centrality scores. (Verma, 2019[12]) presented a study on GNNs with a method of regularization which was derived from semi-supervised prediction of nodes (Yipeng et al., 2019[7]) used SNAP and NetworkX to analyse the graph structure. They used different models based on graphs and evaluated their performance on this dataset.

Chapter 5 - Proposed Method

To take care of the issue it is pivotal to accurately discover hateful users who convey contemptuous talks and can end up being unsafe for the online society. Most current ways to deal with, describe and recognize abhor discourse, center around content posted in Online Social Networks. They face inadequacies to gather and comment on disdainful behaviour because of the deficiency of OSN content and the subjectivity of hateful behaviour and speech. These restrictions are regularly helped with imperatives that misrepresent the issue, for example, considering just tweets containing despise related words. In this work we address these issues by moving the concentration towards users. This work presents a user driven perspective on hateful speech, making ready for better recognition strategies and comprehension. We utilize a Twitter dataset of 100, 386 users along with 4,972 physically labelled nodes. We see that the users are densely associated, and along these lines continue the detection of the users. We start by analysing the data in the dataset. We use Matplotlib and NetworkX to plot graphs and draw insights from them. We then moved on to using semi-supervised graph network models which showed a huge improvement in performance. We also compare performance of different models in order to have a better insight about which model works best.

Chapter 6 - Experimental Results

6.1 Results of Experimental Data Analysis

Through experimental data analysis, we try to analyse features of nodes which represent hateful users. We also try to compare the neighbours of hateful and normal users. It is assumed that neighbours are likely to show similarities. Thus, we compare proximity of pairs and inspect its effect on the results. This is done in order to increase the robustness of the analysis.

6.1.1 Users in different groups

Table 1. Users in different groups

Type of User	Number of Users
Hateful Users	544
Normal Users	4427
Hateful Neighbours	3471
Normal Neighbours	33564
Banned Users	668
Active Users	99718

As can be seen, the data is highly biased. The number of hateful users is very less as compared to normal users. which is also the case in real time situations but this creates an issue when training models. Almost 90% of the users are normal users. It can also be seen that the proportion of labelled nodes is very less, around 5%. This leads to a drop in precision and recall values even when the accuracy values are high. The solution is to train models which learn neighbourhood characteristics. These findings also indicate that using semi-supervised graph based models will give better results than linear or baseline models.

6.1.2 Date of creation of accounts of hateful users

As can be seen in the figure below, hateful users' accounts were created later than normal users. A possible cause could be banning of such users by Twitter itself according to its policies using internally developed methods for fake account detection. It also observed that even 1-neighbours of hateful users have newer accounts which supports common neighbour behaviour assumption. This gives important insights into the way hateful users behave. It also shows that banning accounts helps only to an extent as hateful users tend to make new fake accounts. The constant ranting is likely to continue even after banning the users. In today's scenario, trolls work the same way.

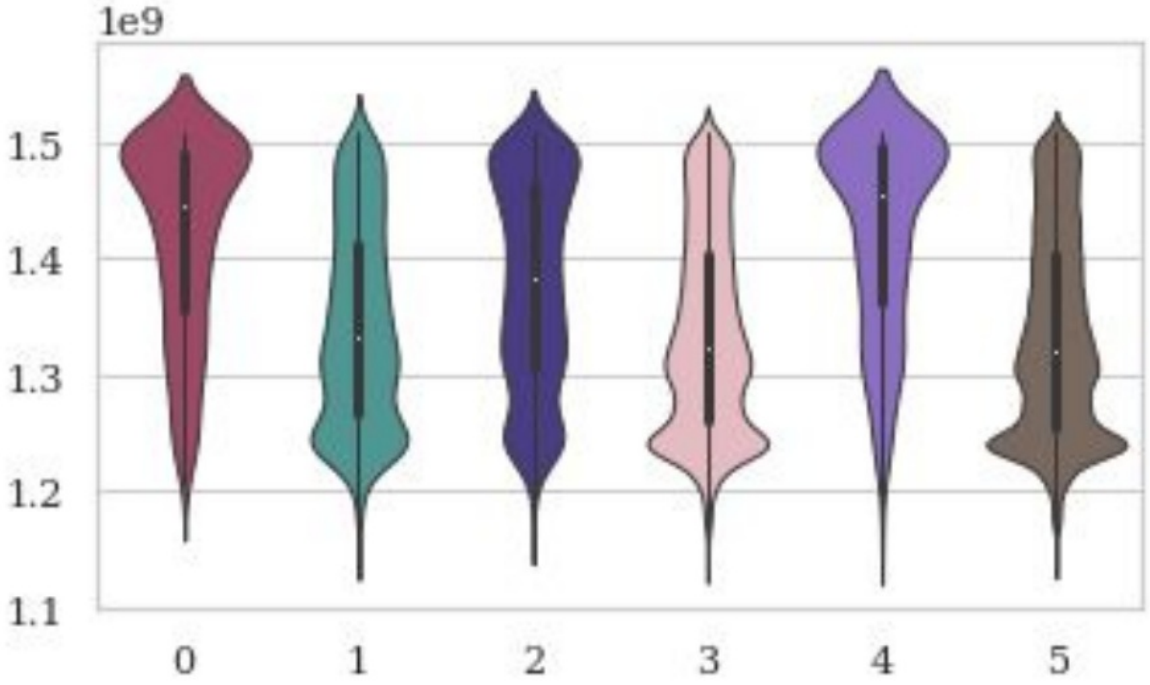


Figure 6.1 : Creation date of users - Left to Right labels: 1. Hateful Users, 2. Normal Users, 3. Hateful Neighbours, 4. Suspended Users, 5. Active Users

Violin Plot for analysis of creation dates of accounts of users.

6.1.3. Behaviour of Hateful Users

We chose metrics like number of followers, tweets, followees and favourite tweets as these reflect the activity of users. These metrics are normalised and shown in the following figure. The results show that hateful users are more active. It can be observed that users practicing hate speech show extreme behaviour in tweeting and liking other user tweets. The analysis of neighbours yields very similar results with the exception of the number of favourite tweets metric.

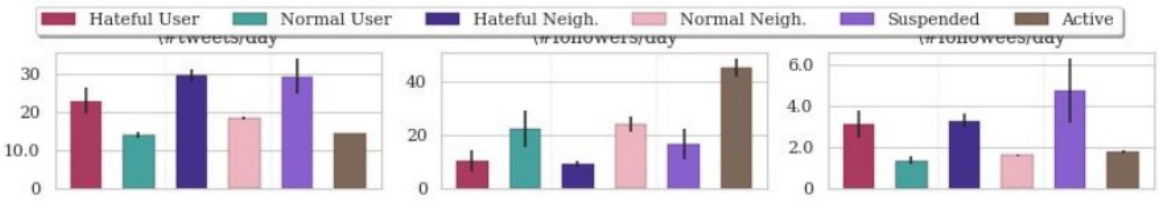


Figure 6.2 : Nature of users - Metrics from left to right - 1. #tweets/day, 2. #followers/day, 3. #followees/day

Bar graph represents average values normalized over time for features like number of tweets, followers and followees.

6.1.4. Comparison of hateful users with spammers

In order to carry this analysis, the metrics used are the ones which are generally used to detect spammers like URLs and hashtags per tweet and followers per followee. We found out that hateful users, on an average, use less hashtags and less URLs. The same holds for their 1-neighbourhood. It was also found that average number of followers per followee are more for normal users. From this, it can be said that hateful users don't have refined content. They don't put efforts in reaching out and are only interested in expressing their anger online. These graphs also point towards the way hateful users prepare their content. It can be said that their content is not organised or systematic. Their content is not engaging as well. **Thus, hateful users are not spammers.** This is all represented through the following figure:

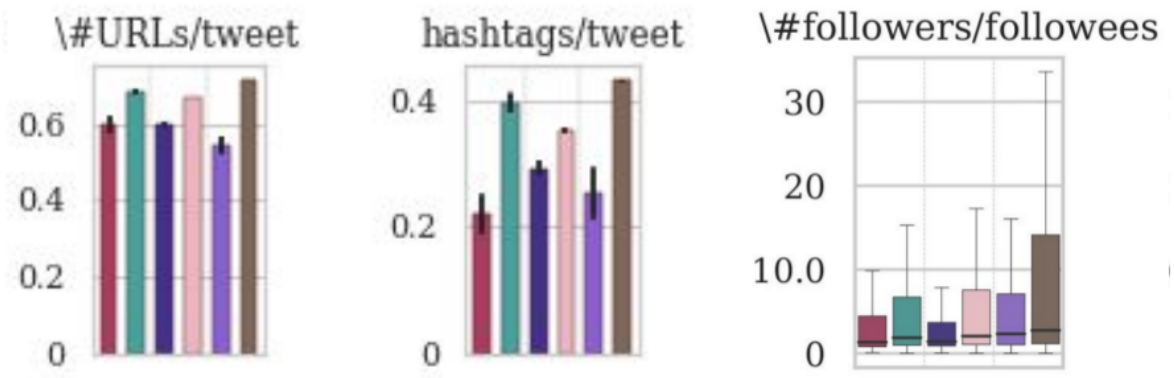


Figure 6.3 : Hateful users v/s spammers - From left to right labels - 1. Hateful Users, 2. Normal Users, 3. Hateful Neighbours

Boxplot and bar the distribution of metrics that indicate spammers.

6.1.5 Centrality of Hateful Users

Different metrics like eigenvector, betweenness and out degree were used to measure how central hateful users were. Although in the real world hate speech is disregarded and people practicing it are usually cast out, we observe in online social networks like twitter, these kinds of users are more outgoing and social. It is confirmed that using internal guidelines, twitter suspends accounts with higher median out degree.

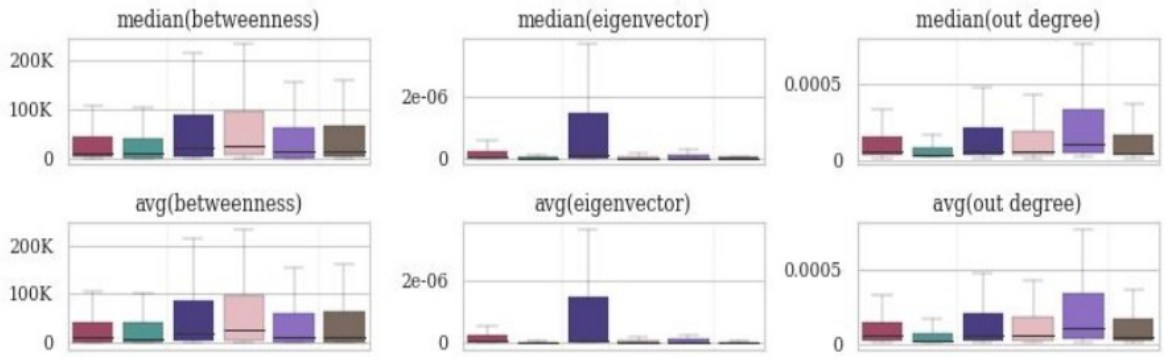


Figure 6.4 : Centrality of Hateful Users - Left to right, top to bottom metrics - 1. median(betweenness), 2. median(eigenvector), 3. median(out degree), 4. avg(betweenness), 5. avg(eigenvector), 6. avg(out degree)

Box-plots showing average and measures for centrality of users.

6.1.6 Vocabulary used in hate speech

We realize that contrary to assumptions made about generic hate speech containing extreme emotions like hate, terrorism and anger, hateful users portray more of emotional vocabulary in their speech. Much of the hate speech is driven by users who portray positive emotions more than normal users which is contradictory in nature and thus this brings to our notice that just NLP analysis of tweets/hate speech in online social networks is ineffective in categorising hateful users. NLP analysis still fails to capture instances of sarcasm, code language and slang by users practicing hate speech and thus strengthening our motivation for a user based approach.

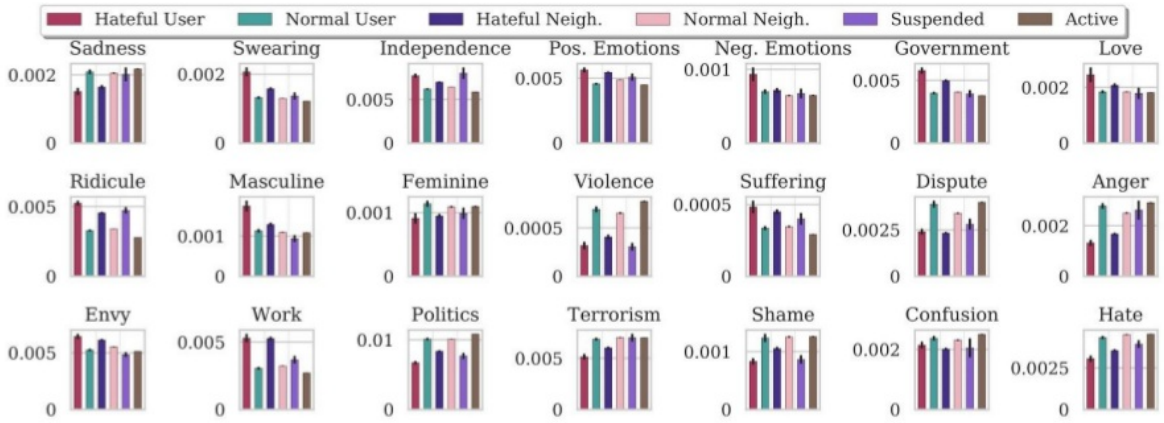


Figure 6.5 : Vocabulary of hateful users

Bar graph representing average values of occurrence of different types of emotions per tweet. It can be observed that these categories cover a bunch of different emotions and give insights on vocabulary of user.s

6.1.7 Sentiment of Hateful Users

We observe that sentences composed by these users are progressively negative, and are less subjective. The neighbors of these users are progressively negative as well but not less emotional. We additionally find out the circulation of profane words per tweet in hateful users. We found out that suspended users, hateful clients and their neighbors utilize increasingly profane words per tweet, likewise affirming the outcomes from the investigation with Empath. It can be shown through the following figure:

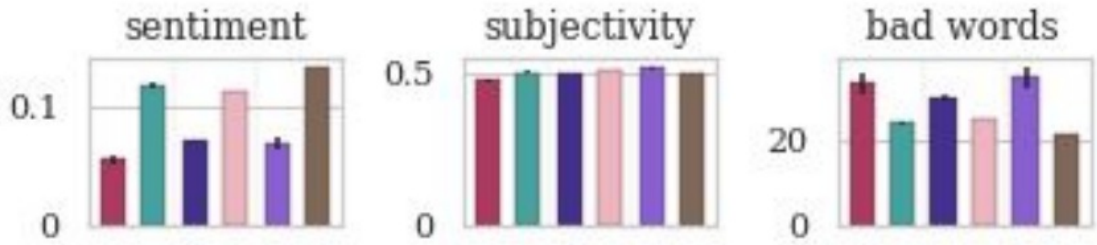


Figure 6.6 : Sentiment of hateful users - From left to right labels - 1. Hateful Users, 2. Normal Users, 3. Hateful Neighbours, 4. Suspended Users, 5. Active Users

Bar graphs showing distribution of different characteristics used to classify sentiment of a tweet.

6.1.8 Connectivity of Hateful Users

From our analysis it is observed that the retweet density among hateful users is extremely concentrated, The same results can be observed among suspended accounts which strengthens the fact that most suspended accounts were once hateful users, due to their high correlation in connectivity.We leverage this finding further on by application of graph based machine learning models in detecting these users.

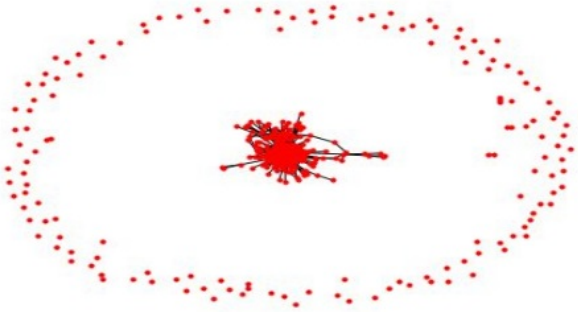


Figure 6.7 : Plot of hateful users and how they are connected. Made using NetworkX library.

6.2 Models Used For Training

6.2.1 GraphSAGE

We experimented with a semi-supervised classifier that used GraphSAGE operators for hateful user node classification with Pytorch geometric. GraphSAGE is a general inductive framework that makes use of node features and graph information to efficiently generate node embeddings for previously unseen data. With these generated embeddings, we conducted a binary classification on each node for hateful user detection. The main steps were creating embedding using information from the neighbours and aggregating it after sampling, update and normalization. We used a multilayer perceptron.

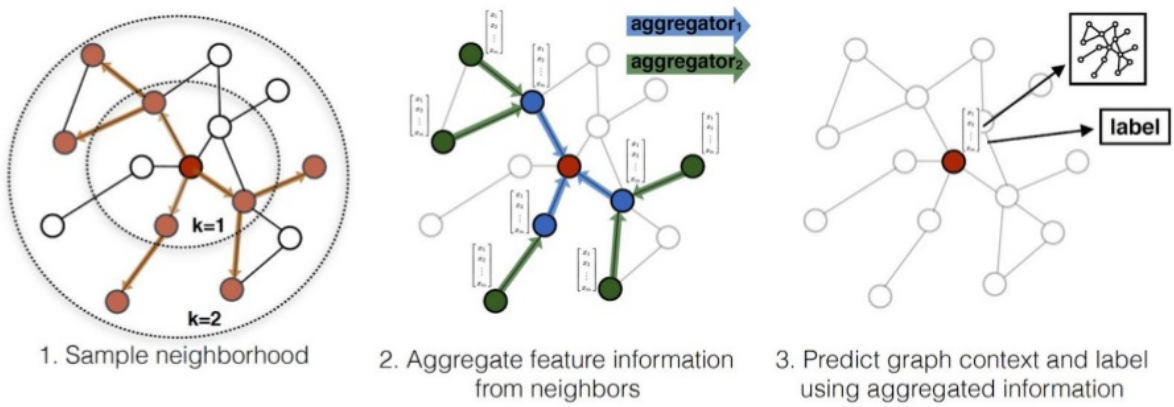


Figure 6.8 : Steps in GraphSAGE [13]

.

After preprocessing our data for the model and training using GraphSAGE, the epoch-wise distribution of loss looks as below:

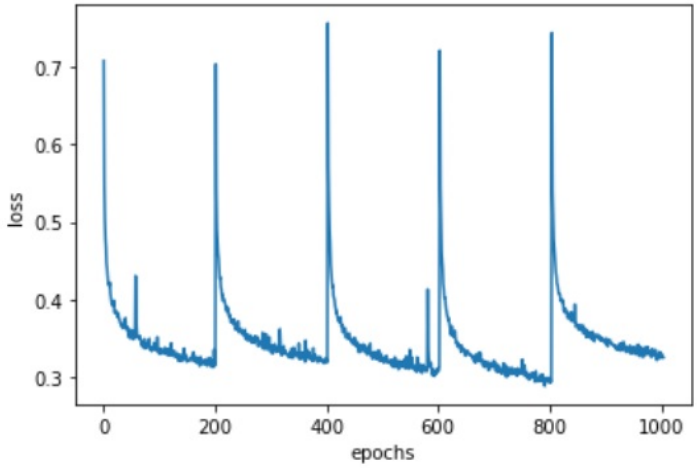


Figure 6.9 : GraphSAGE loss v/s epochs

6.2.2 Graph Convolutional Network(GCN)

GCN are similar in concept to a convolutional neural network (CNN), that is they have multiple hidden layers where each hidden layer can be composed as Y(i) = p(Y(i-1),A) where A is the adjacency matrix and Y(0) is our feature matrix.We consider this model as a baseline model of performance for the further variants of GCN we use.

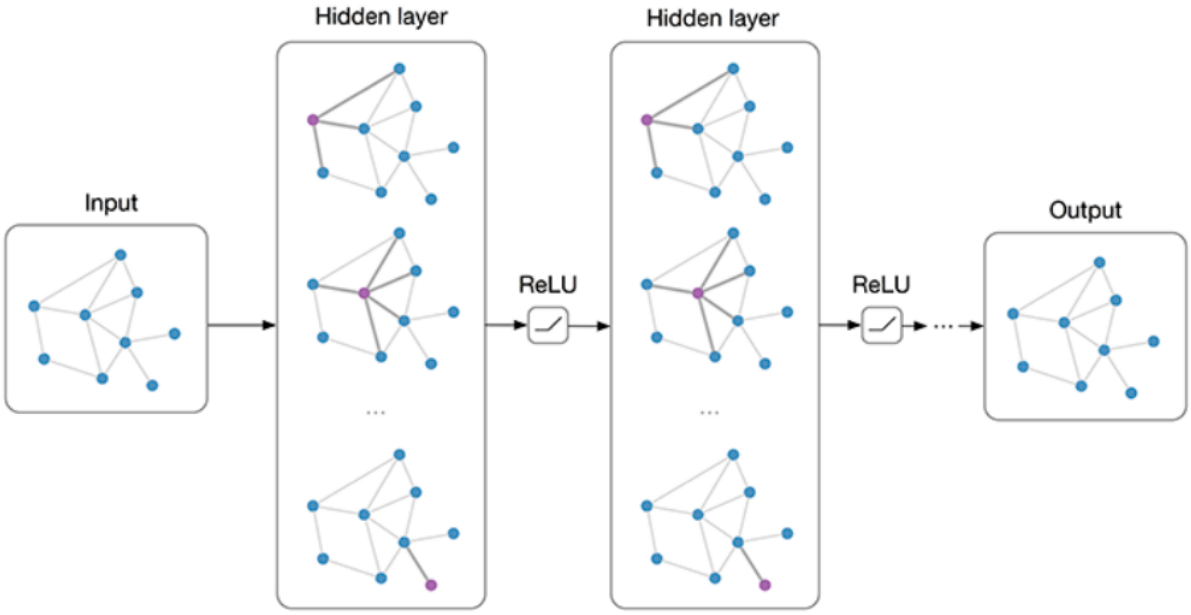


Figure 6.10 : Basic GCN architecture [14]

After preprocessing our data for the model and training using GCN, the epoch-wise distribution of accuracy and loss looks as below:

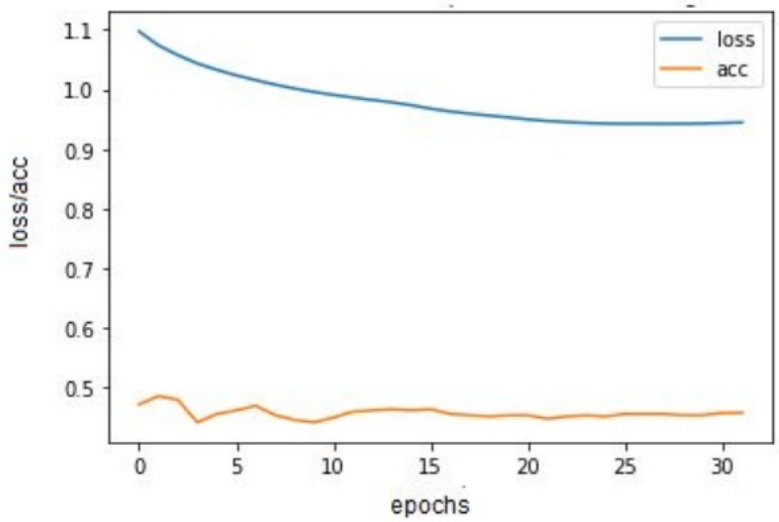


Fig 6.11 : GCN accuracy and loss v/s epochs

6.2.3 Hierarchical Graph Convolutional Networks(H-GCN)

We use HGCN to combat issues in the previous model such as shallow neighborhood aggregation and absence of graph pooling mechanism, as they are significant in obtaining global information of the twitter retweet network.

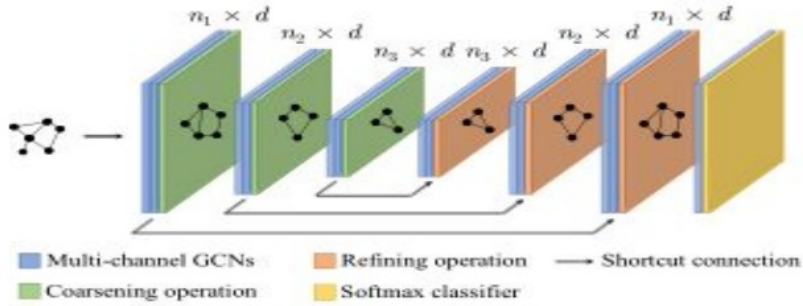


Figure 6.12 : HGCN architecture [16]

HGCN uses the pooling concept by aggregating similar nodes. The receptive field for each node is enlarged by the implemented coarsening procedure and hence global information of the graph structure is not lost.

After preprocessing our data for the model and training using HGCN, the epoch-wise distribution of

accuracy and loss looks as below:

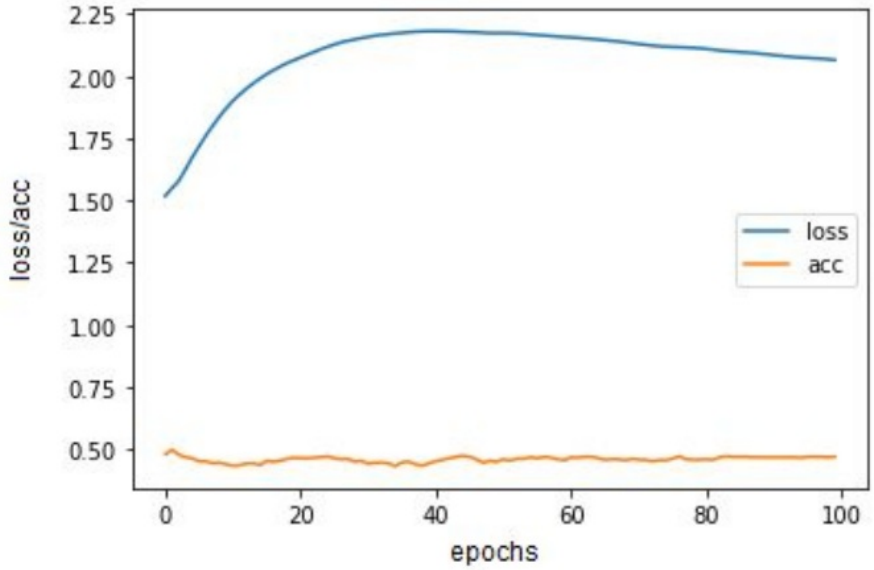


Figure 6.13 : HGCN loss and accuracy v/s epochs

6.2.4 Bayesian Graph Convolutional Neural Networks(BGCN)

Although graph convolutional neural networks(GCNNS) have impressive performance and address node and graph classification, the current implementations of these networks have limits in incorporating uncertainty in the graph structure. GCNNS portray a good depiction of the relationships in edges in the graph structure, they often are modelled from real world noisy data and flawed assumptions.Important edges denoting very strong relationships between nodes may be missing and spurious edges might be included.

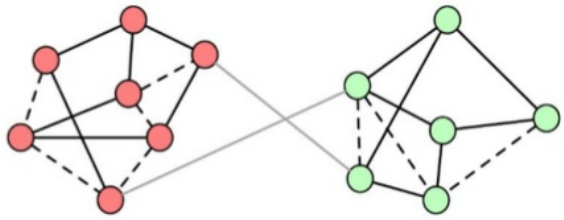
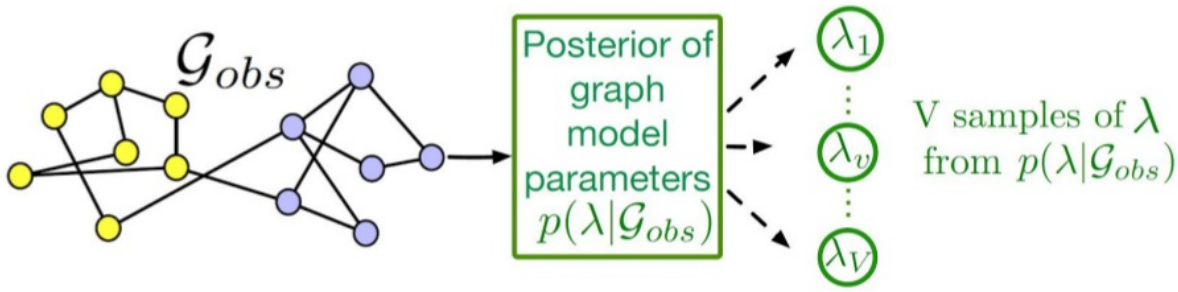


Figure 6.14 : Graph edges in non-bayesian approach [17]

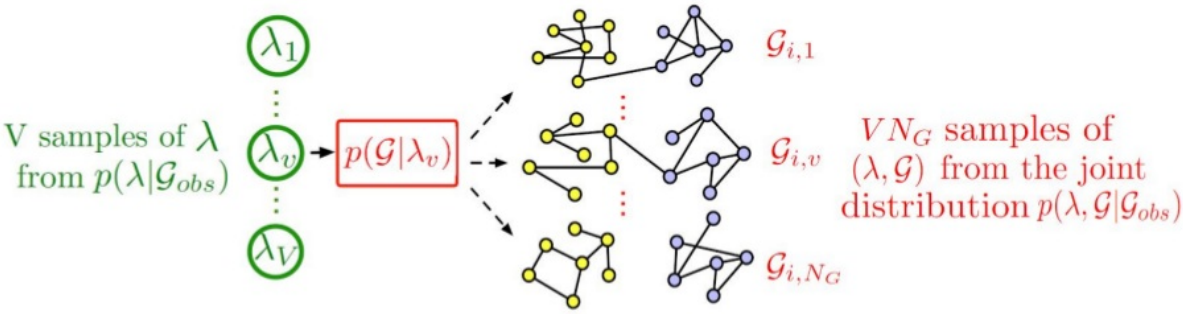
To avoid the above, we adopted a model based on Bayesian approach which views the graph as a realization from a parametric family of random graphs. The model then target-inference of the joint

posterior of parameters of the random graphs and the node(graph) labels.It develops an iterative-learning procedure for the ease of mixed membership stochastic block models.[17]

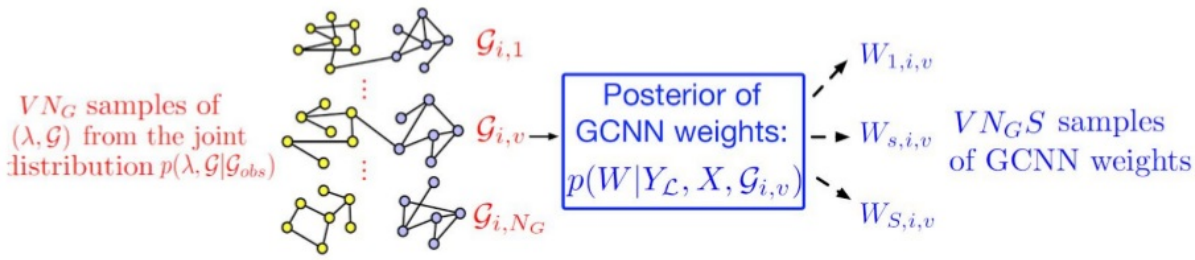
Steps of a BGCN depicted in a visual form



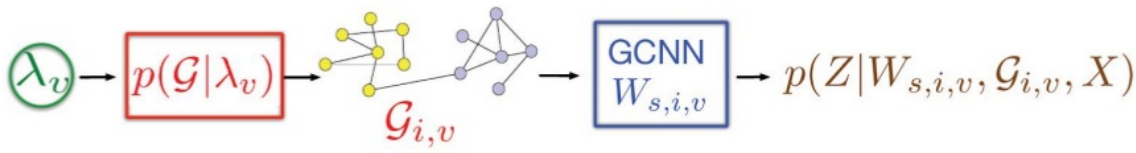
$$p(\mathbf{Z}|\mathbf{Y}_{\mathcal{L}}, \mathbf{X}, \mathcal{G}_{obs}) = \int p(\mathbf{Z}|W, \mathcal{G}, \mathbf{X}) p(W|\mathbf{Y}_{\mathcal{L}}, \mathbf{X}, \mathcal{G}) p(\mathcal{G}|\lambda) p(\lambda|\mathcal{G}_{obs})\, dW\, d\mathcal{G}\, d\lambda\,.$$



$$p(\mathbf{Z}|\mathbf{Y}_{\mathcal{L}}, \mathbf{X}, \mathcal{G}_{obs}) = \int p(\mathbf{Z}|W, \mathcal{G}, \mathbf{X}) p(W|\mathbf{Y}_{\mathcal{L}}, \mathbf{X}, \mathcal{G}) p(\mathcal{G}|\lambda) p(\lambda|\mathcal{G}_{obs})\, dW\, d\mathcal{G}\, d\lambda\,.$$



$$p(\mathbf{Z}|\mathbf{Y}_{\mathcal{L}}, \mathbf{X}, \mathcal{G}_{obs}) = \int p(\mathbf{Z}|W, \mathcal{G}, \mathbf{X}) p(W|\mathbf{Y}_{\mathcal{L}}, \mathbf{X}, \mathcal{G}) p(\mathcal{G}|\lambda) p(\lambda|\mathcal{G}_{obs})\, dW\, d\mathcal{G}\, d\lambda\,.$$



$$p(\mathbf{Z}|\mathbf{Y}_{\mathcal{L}}, \mathbf{X}, \mathcal{G}_{obs}) = \int p(\mathbf{Z}|W, \mathcal{G}, \mathbf{X}) p(W|\mathbf{Y}_{\mathcal{L}}, \mathbf{X}, \mathcal{G}) p(\mathcal{G}|\lambda) p(\lambda|\mathcal{G}_{obs})\, dW\, d\mathcal{G}\, d\lambda\,,$$

$$\approx \frac{1}{V} \sum_{v=1}^{V} \frac{1}{N_G S} \sum_{i=1}^{N_G} \sum_{s=1}^{S} p(\mathbf{Z}|W_{s,i,v}, \mathcal{G}_{i,v}, \mathbf{X})\,.$$

[17]

After preprocessing our data for the model and training using BGCN, the epoch-wise distribution of accuracy and loss looks as below:

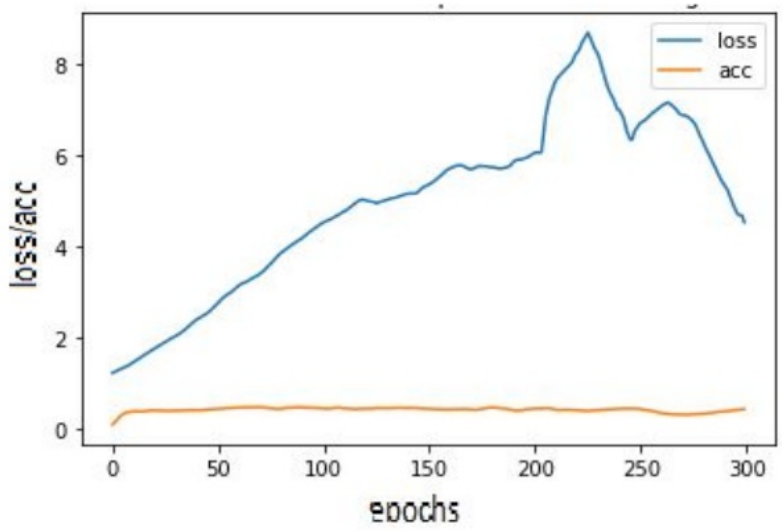


Figure 6.15 BGCN loss and accuracy vs epochs

Results from various models look as below:

Table 2. Model results

Model	Loss	Accuracy
GCN	0.94714	0.61051
HGCN	2.14972	0.55496
BGCN	**0.31735**	**0.93333**
GraphSAGE	0.3411	0.8489

Chapter 7 - Conclusion

We experimented with different models to perform classification of nodes of the retweet graph. We aimed at classifying users as hateful or normal. Our biggest hurdle was that we had only limited data which was labelled, that is, only 5% of our total data was labelled. We used a bunch of semi supervised models which were pretrained on datasets like Cora or Pubmed and were based on graph classification. These models included GraphSAGE and different types of GCNs. We want to say in conclusion that graph based networks work better than baseline or linear models due to their ability to store information from neighbours and learn the structure of graphs. Graph based neural networks help in overcoming limitations which arise with NLP methods to detect hate speech in online social networks as hate speech can be composed of variations of random emotions which can cause conflicting results in hate speech detection. In contrast, graph based networks work better than linear models and NLP methods as they leverage the graphical structure properties of online social networks.

References

[1] https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

[2] https://www.geeksforgeeks.org/kde-plot-visualization-with-pandas-and-seaborn/

[3] https://en.wikipedia.org/wiki/Centrality

[4] https://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors

[5] https://www.techopedia.com/definition/12937/network-connectivity

[6] https://arxiv.org/pdf/1803.08977.pdf

[7] http://web.stanford.edu/class/cs224w/project/26424341.pdf

[8] David R Bild, Yue Liu, Robert P Dick, Z Morley Mao, and Dan S Wallach. 2015. Aggregate characterization of user behavior in twitter and analysis of the retweet graph. ACM Transactions on Internet Technology (TOIT), 15(1):4.

[9] Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgilio AF Almeida, and Wagner Meira Jr. 2017. " like sheep among wolves": Characterizing hateful users on twitter. arXiv preprint arXiv:1801.00317.

[10] Min-Chul Yang, Jung-Tae Lee, Seung-Wook Lee, and Hae-Chang Rim. 2012. Finding interesting posts on twitter based on retweet graph analysis. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pages 1073–1074. ACM.

[11] Phi Vu Tran. 2018. Learning to make predictions on graphs with autoencoders. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pages 237–245. IEEE.

[12] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, David Lopez-Paz, and Yoshua Bengio. 2018. Manifold mixup: Better representations by interpolating hidden states. arXiv preprint arXiv:1806.05236.

[13] https://towardsdatascience.com/an-intuitive-explanation-of-graphsage-6df9437ee64f

[14] https://tkipf.github.io/graph-convolutional-networks/

[15] https://towardsdatascience.com/how-to-do-deep-learning-on-graphs-with-graph-convolutional-networks-7d2250723780

[16] http://snap.stanford.edu/hgcn/

[17] https://github.com/huawei-noah/BGCN

Project_Report_3.pdf

ORIGINALITY REPORT

5%

SIMILARITY INDEX

PRIMARY SOURCES

1	web.stanford.edu Internet	73 words — 2%
2	eprints.soton.ac.uk Internet	41 words — 1%
3	manualzz.com Internet	21 words — < 1%
4	www.geeksforgeeks.org Internet	18 words — < 1%
5	www.arc.gov Internet	13 words — < 1%
6	"Advances in Knowledge Discovery and Data Mining", Springer Science and Business Media LLC, 2020 Crossref	13 words — < 1%
7	www.scribd.com Internet	11 words — < 1%
8	www.intechopen.com Internet	10 words — < 1%
9	vuir.vu.edu.au Internet	9 words — < 1%
10	docplayer.net Internet	8 words — < 1%

11 Binny Mathew, Ritam Dutt, Pawan Goyal, Animesh Mukherjee. "Spread of Hate Speech in Online Social Media", Proceedings of the 10th ACM Conference on Web Science - WebSci '19, 2019
Crossref

6 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	ON		