Consumption of Hate Speech on Twitter: A Topical Approach to Capture Networks of Hateful Users

Sameer Gupta[a], Seema Nagar[b], Amit Anil Nanavati[c], Kuntal Dey[d], Ferdous Ahmed Barbhuiya[e] and Sougata Mukherjea[f]

[a]National Institute of Technology, Kurukshetra
[b]IBM India Pvt. Ltd. Bangalore, India
[c]IBM India Pvt. Ltd. New Delhi, India
[d]Accenture Technology Labs, Bangalore
[e]Indian Institute of Information Technology, Guwahati
[f]IBM India Pvt. Ltd. New Delhi, India

Abstract

In this paper, we attempt to track the dissemination of hate speech on Twitter. We argue that hate is not a blanket category but exists across multiple topics. We use topic modelling to unearth the latent topics in tweets and an ensemble classification model to capture various nuances of hate speech. We further validate our approach by manually annotating 4,720 tweets. On analysing the mechanisms of hate speech dissemination, we find that hateful tweets garner 2.4 times more retweets than non-hateful tweets. Further, the retweet network allows us to define topical inflow and outflow vectors which we use to classify users as *originators*, *propagators*, and *constrictors*. We then examine how these users take part in the dissemination of topical information and observe considerable differences in how users associate with hateful and non-hateful topics. Furthermore, the retweet network enables us to analyse the structure of the strongest connected components for different topics. We find some topics associate strongly with hate and users associated with these topics have high degree centrality, are densely connected, and have a large strongly connected core.

Keywords

Hate detection, Hate spread, Topic modelling

1. Introduction

Hate Speech has taken a catalysing form in moulding the opinions of people on Online Social Networks (OSN). It has also incited real life violence with its spread like wildfire among users [1, 2, 3, 4]. The nature of spread of hateful content is approached by many within the community. Ribeiro et. al. [4] find that contrary to the general assumption that hateful users are alone, they are actually central in their social network. Mathew et. al. [2] present a study of hate and counter speech accounts on Twitter. They further study the spread of hateful content on Gab (gab.com), which hosts a large amount of hateful content [3].

ROMCIR 2021: Workshop on Reducing Online Misinformation through Credible Information Retrieval, held as part of ECIR 2021: the 43rd European Conference on Information Retrieval, March 28 – April 1, 2021, Lucca, Italy (Online Event)
sameer.lego@gmail.com (S. Gupta); senagar3@in.ibm.com (S. Nagar); namit@in.ibm.com (A. A. Nanavati); kuntal.dey@accenture.com (K. Dey); ferdousa@gmail.com (F. A. Barbhuiya); smukherj@in.ibm.com (S. Mukherjea)

Prior works treat hate as a blanket topic for modelling spread. However, there are various topics in the universe of Twitter such as religion, politics, sports, among others and some of these topics are associated with the dissemination of a larger proportion of hateful content as compared to others . Hence, we model the spread of hate under the purview of the dissemination of topical information.

The characterization of users as hateful or non-hateful is prone to errors due to the subjective and contextual nature of hate [5, 6]. Hence, by incorporating a topic-model combined with a user-centric analysis of hate speech, we are able to view a top-down snapshot of the networks of hate. We use an ensemble classification model to examine the fraction of tweets that are classified as hateful for each user. Further, we validate our approach by manually annotating 4,720 tweets as hateful or not. Moreover, we study the different roles played by users in this dissemination. Using the retweet network, we define topical inflow and outflow vectors which we use to classify users as *propagators*, *constrictors* or *originators*. Furthermore, by exploring the topical subgraphs, we gain insights into the core structures of these topical communities. This provides an understanding of the characteristics of intra-topic information flow.

We are neither aiming to outperform the state of the art, nor do we wish to solely present a state-of-the-art machine learning approach for hate speech detection. Instead, we are modifying existing approaches as a base for downstream our research tasks. We have shown the efficacy of analysing hate and networks of hateful users through a topical lens on online social media platforms. Further, due to the large size of the dataset, we have utilised simple machine learning models to improve the classification speed, while also achieving decent performance on the classification task.

To summarise, we make the following contributions in this paper, a) propose an ensemble based learning approach for hate detection, b) manually annotate 4,720 tweets as hateful or not, c) analyse hateful content present in different topics, d) propose a method to compute *propagators*, *originators* and *constrictors* for a topic and e) empirically demonstrate on a very large Twitter dataset that users associating with hateful topics are densely connected, form a dense strongly connected core, and are more central in the retweet network.

2. Proposed Approach

We present a detailed description of the proposed approach in this section. As the dataset we use does not have labels for tweets, we first explain how we build an ensemble classifier for hate detection which is used to label the data. We then explain how we model spread of hate along multiple topics.

2.1. Ensemble based learning approach for hate classification

Ensemble based classification models have performed better than individual classification models [7]. We utilize a variety of data available for hate detection that captures the nuances of hate speech such as toxicity, obscenity, threats, insults, identity hate, racism and sexism.

Two classification models, A and B, are trained independently on two separate datasets. We take the mean of the predictions as the final classification score. Both models follow Wang et. al.[8]'s philosophy that the feature vector is computed using Naive Bayes log count ratios,

bi-grams and unigrams. However, we use logistic regression instead of support vector machine for classification as it performs better in our experiments. Moreover, we manually annotate 4,720 tweets as hateful or not to as a validation set for our approach.

2.2. Hate Spread Modelling

We model the spread of hate by applying the following steps: a) topic modelling to detect latent topics, b) associating topics with hate, and c) classifying users based on their contribution to topics.

Topic Modelling: We create a corpus of N documents for N users by combining all the tweets of each user into one document. We then train a topic model on this corpus using Latent Dirichlet Allocation (LDA), obtaining a topic vector $T(i)_{(1\times|t|)}$ for each user i and set of topics t. Here, $|t|$ is the number of topics.

Associating Topics with Hate: Each topic has different proportions of hateful content. We quantify the association of a topic with hate using the following steps:

1. For each user, classify their tweets as hateful or non-hateful and obtain $H(i)$ as the fraction of tweets that are hateful. This represents the association of a user i with hate.
2. Obtain the topic vector $T(i)_{(1\times|t|)}$ for each user i using the topic model,
3. For each topic $t_a \in t$, obtain the set of users M, having fractional interest $T(i)(t_a)$ greater than a threshold τ,
4. Compute $hate\_assoc(t_a) = 1/|M|(\sum_{i=1}^{M} H(i))$ where $|M|$ is the size of the set.

Defining Inflow and Outflow: Intuitively, the inflow topic vectors are the topics a user is exposed to originating from the content generated(tweets, retweets, quotes) by users they interact with. Similarly, outflow topic vectors constitute the topics in the content generated by a user.

The inflow topic vector is defined as follows:
For each user i:

1. From the retweet-induced graph, select the set of users J who have been retweeted by user i,
2. $I(i)_{(1\times|t|)} = 1/|J|(\sum_{j=1}^{J} T(j)_{(1\times|t|)})$ where j is a user in the set, $|J|$ is the size of the set, and $|t|$ denotes the number of topics.

Similarly, the topic vector $T(i)_{(1\times|t|)}$ represents the distribution of topics in the content generated by a user i. Thus, we use the outflow topic vector as the topic vector for the user.

Defining Types of Influential Users: We define three categories of users based on their role in the dissemination of a topic, namely, *originators*, *propagators* and *constrictors* using the approach mentioned in Algorithm 1.

Users associate differently with various topics. Thus, we adjust the m_{max} and m_{min} values, which are defined as the maximum and minimum value of the difference between the outflow

Algorithm 1 Classifying a user into one of the three categories with respect to each topic

Input: Set of users U, Topic vector $T(i)_{(1\times|t|)}$ and Inflow vector $I(i)_{(1\times|t|)}$ $\forall i \in U$
Output: Category of the user for each topic

for each topic $t_a \in t$ $\forall a \in |t|$ **do**
  $m_{max} \leftarrow \max(T(i)(t_a) - I(i)(t_a))$ $\forall i \in U$
  $m_{min} \leftarrow \min(T(i)(t_a) - I(i)(t_a))$ $\forall i \in U$
  for each user i **do**
    $d \leftarrow T(i)(t_a) - I(i)(t_a)$
    if $d \geq 0.5 \times m_{max}$ **then**
      categorise i as an *originator*
    end if
    if $0.1 \times m_{max} \leq d < 0.5 \times m_{max}$ **then**
      categorise i as a *propagator*
    end if
    if $d \leq 0.1 \times m_{min}$ **then**
      categorise i as a *constrictor*
    end if
  end for
end for

and the inflow vector, respectively. The parameter d represents the net topical outflow for a user.

We define *originators* as users whose d value is higher than at least half (50%) of the maximum net topical outflow across all the users (m_{max}). For a particular topic t_a and user i, as the sum of the outflow vector $\sum_{a=1}^{|t|} T(i)(t_a) = 1$, a high value of d implies a high value of the fractional interest $T(i)(t_a)$. This suggests that the user's content mostly contains information discussing the particular topic.

Similarly, *propagators* are users whose d value is less than those of originators but greater than at least 10% of m_{max}. The lower limit of 10% is empirically set to account for neutral users whose d value is quite low.

Further, we define *constrictors* as users who restrict the flow of topical information. Thus, the d value for these users is always negative and lower than a threshold of 10% of the minimum net topical outflow across all the users (m_{min}). This threshold serves the same purpose as explained above.

As per our observations, the values of m_{max} were always between 0 and +1 and the values of m_{min} were always between 0 and −1.

3. Experiments and Results

3.1. Dataset, Preprocessing, and Annotation

Dataset: We use the dataset provided by [4]. This dataset contains 200 most recent tweets of 100,386 users, totaling to about $19M$ tweets. Furthermore, each tweet is categorized as an original tweet ($10.7M$), retweet ($7.23M$) or a quote ($1.6M$). A retweet induced graph with 2,286,592 directed edges is also provided. The retweet-induced graph is a directed graph $G = (V, E)$ where each node $u \in V$ represents a user in Twitter, and each edge $(u1, u2) \in E$ represents a user $u1$ retweeting user $u2$. As the influence flows in the opposite direction of retweets, we work on the graph with inverted edges. Intuitively, given that a lot of people retweet user u_i and u_i retweets nobody, u_i may still be a central and influential node.

Out of the 100,386 users, labels (hateful or normal) are available for 4,972 users, out of which 544 users are labelled as hateful and the rest as normal. This dataset does not have labels for the tweet content.

Though the original dataset does not have labels for the tweet content, we use it for two reasons, a) it is the largest collection of users marked as hateful and b) it gives us an opportunity to examine the retweet-induced graph structure.

Preprocessing: We first convert tweet text to lower case and then remove the stop words. We also remove numbers, punctuation, special characters and emoticons, HTTP links, user mentions such as @{user} and RT{user}, trailing spaces, and condense instances of trailing letters to a single occurrence, such as lmaooo to lmao and lolll to lol.

Annotation: As the dataset we use does not have labels for the tweets, we manually annotate the tweets as hateful or not to create a validation set for our ensemble classification model. Annotating $19M$ tweets is very expensive and time consuming process. Hence, we annotate a subset of tweets. We pick the 10 original tweets produced by each of the 544 users labelled as hateful in the dataset for two main reasons. Firstly, annotating the tweets of hateful users ensures that we get an adequate number of hateful tweets. Secondly, this annotated set serves as a validation ground truth for downstream tasks in our research. After preprocessing the tweets, we were left with 4,720 original tweets.

The tweets were annotated by a group of 4 independent annotators whose primary language is English. The Inter-Annotator Agreement (IAA) score (Cohen κ) was 0.87. A high κ score was obtained due to a singular class - hateful or not.

3.2. Ensemble Based Hate Classification

We use an ensemble based classification model to label the dataset we use for our analysis. We use two separate models, A and B for our ensemble model. Model A learns from a dataset of toxic comments, while model B learns a dataset of hate speech against immigrants and women. Both the models consist of a logistic regression classifier with Naive Bayes log count ratio as features. We implement the models using the scikit-learn[1] library. Further, we split both the datasets into a 80:20 train-test split. We empirically tune the parameters for the model to maximize the F1 score.

[1] https://scikit-learn.org/stable/

Model A is borrowed from a kernel from Kaggle posted for the competition "Toxic Comment Classification Challenge"[2]. The dataset contains $160K$ labelled (hateful or non-hateful) comments. We add tf-idf with unigrams and bigrams, set minimum document frequency to 3[3] and maximum document frequency to 0.9[4]. The model achieves an accuracy of 95% with a F1 score of 0.8 on the test data. This model achieves an accuracy of 72% with a F1 score of 0.69 on our annotated dataset (validation set).

Model B follows the same principal as model A. However, we modify the feature space by not specifying the minimum and maximum document frequency. Model B is trained on a dataset from a Semeval challenge called "Multilingual detection of hate speech against immigrants and women in Twitter (hatEval)"[5]. This dataset consists of $9,000$ labelled tweets. The model achieves accuracy of 80% with a F1 score of 0.7 on the test data. This model achieves an accuracy of 78% with a F1 score of 0.73 on our annotated dataset (validation set).

The final predictions are performed separately using both the models and the mean confidence score is used to predict a class (hateful or non-hateful). The ensemble model achieves an accuracy of 82% with a F1 score of 0.75 on the annotated dataset (validation set).

We then proceed to classify our whole corpus of $19M$ tweets using our ensemble model. From the classification results, we infer that out of $1.6M$ quotes, 12.4% are hateful, out of $7.23M$ retweets, 8.36% are hateful and out of $10.7M$ original tweets, 7% are hateful. This suggests that quoting and retweeting are popular mechanisms of propagating hateful content. On comparing the influence of tweets we find that, on an average, hateful tweets garner 2.4 times more retweets than non-hateful tweets.

3.3. Topic Modelling Results

Topic Modelling: We analyse the temporal spread of tweets to ascertain if they represent the user's interests adequately. We find that 80% users have a temporal spread greater than a week, leading us to the conclusion that for most users, the interest in topics is not transitory in nature.

We use MALLET[6] to train a LDA based topic model on the corpus (N documents for the N users). We vary the number of topics $|t|$ from 10 to 80 as 10, 20, 40, 80. We empirically tune the topic density parameter (α) and the number of topics to maximize the coherence score and minimize the overlap between topics. We chose the number of topics to be 40 and α to 0.01. A smaller value of α leads to more concentrated topical distributions. We obtain topic vectors $T(i)$ for each user i, represented as a fraction of interest in each of the $t_a \in t \; \forall a \in |t|$ topics. The topic vector matrix is a $N \times |t|$ matrix, where N is the number of users and $|t|$ is the number of topics.

Associating hate with each topic: We calculate the $hate\_assoc(t_a)$ as described in Section 2 for each $t_a \in t$. We set the threshold τ to 0.5, which implies that the fractional interest $T(i)(t_a)$ of a user i for a particular topic t_a is more than 50%. To validate our approach, we first find the top 10 topics discussed by the 544 labelled ground truth hateful users. For this, we find the topic t_a that has the max $T(i)(t_a) \; \forall a \in |t|$, for each user i in the set. The top 10 topics are ascertained by

[2]https://www.kaggle.com/jhoward/nb-svm-strong-linear-baseline

[3]This implies removing the terms which appear in less than 3 documents

[4]This implies removing the terms which appear in more than 90% of the documents

[5]https://competitions.codalab.org/competitions/19935

[6]http://mallet.cs.umass.edu/

sorting the number of occurrences of these t_a in a descending order. Finally, we analyse the overlap of these top 10 topics with the top 10 hateful topics calculated based on descending $hate\_assoc(t_a)$ values.

We find that 7 out of the 10 topics are common and the top 5 topics are same for both the approaches. Hence, we use the top 5 topics as hateful topics for further analysis. Further, we find that there are considerable differences in the top words for the top 5 hateful topics and the top 5 least hateful topics as shown in Figure 1.

Figure 1: Words associated with top 5 hateful topics in red and non-hateful topics in blue.

3.4. Spread Results

Inflow and Outflow Results: Figure 2 shows the average inflow and outflow distributions for the top 5 hateful and non-hateful topics. We observe that, at lower values of contribution, there are more users that associate with non-hateful topics. However, as the values of contribution increase, very few users associate with non-hateful topics, whereas, a considerably higher number of people associate with hateful topics.

Node Influence Metrics: Figure 3 shows the distribution of *originators*, *propagators* and *constrictors* for each topic. We observe that there are a considerably higher number of propagators for hateful topics as compared to non-hateful topics.

Static Network Structure Related Properties: For each topic $t_a \in t \; \forall a \in |t|$, we create a topic graph in following manner:

- Obtain the set of users M, having fractional interest $T(i)(t_a)$ greater than a threshold $\tau = 0.5$ (we set this value empirically to include only those users who associate strongly with a particular topic)
- For each of these users u_x, obtain their ego network, containing edges (u_x, u_y) and (u_y, u_x) from retweet-induced graph $G = (V, E)$
- Take the union of the ego networks to create a topic graph

We analyse centrality metrics, average path lengths, and the structure of strongly connected components (SCCs) for the topic graphs. From Figure 4, we observe that users associated with hateful topics have 2.67 times the average degree centrality that non-hateful topics. Moreover, we find that non-hateful topics have 1.33 times the average path lengths of hateful topics. This

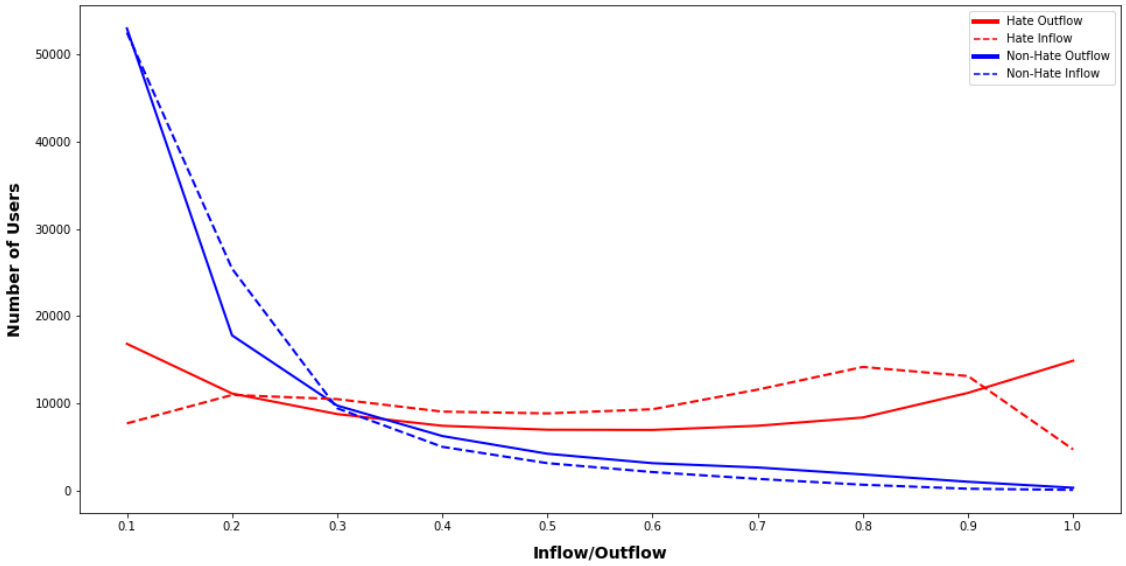

Figure 2: Distribution of inflow and outflow values for the top 5 hateful and non-hateful topics.

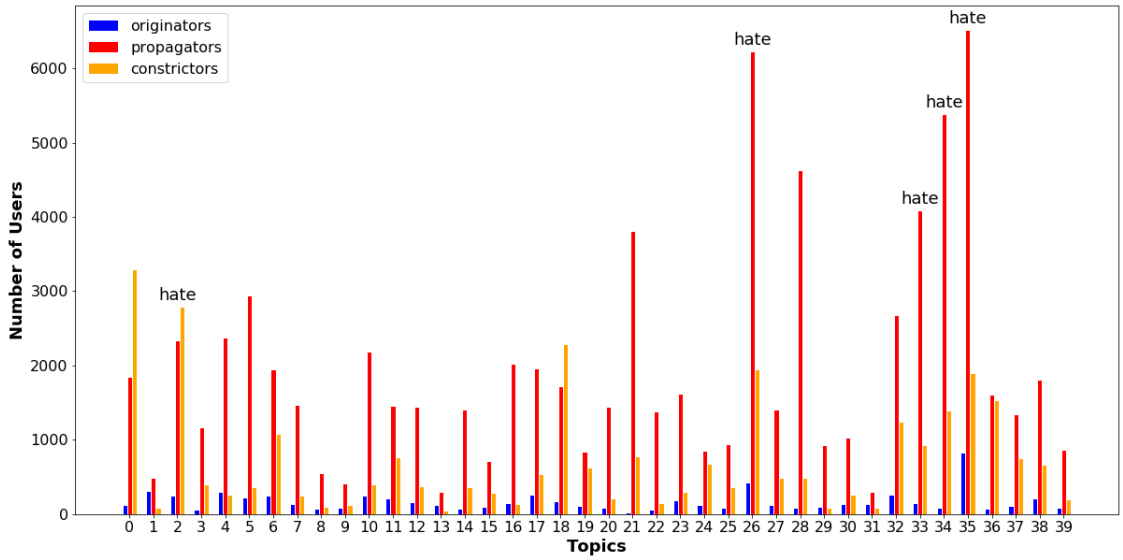

Figure 3: Topic wise distribution of users across the three categories.

indicates a more central position in the network and supplements the results of the study conducted by Ribeiro et. al. [4].

From Figure 5, we observe that the SCC of a hateful topic graph has a strong core while it is relatively sparse for the non-hateful topic. Hateful topics also have more users in the largest SCC; the mean number of users for the hateful topics is 11,875 compared to 4,050 for non-hateful topics. Moreover, we observe that hateful topics correlate with denser networks; the

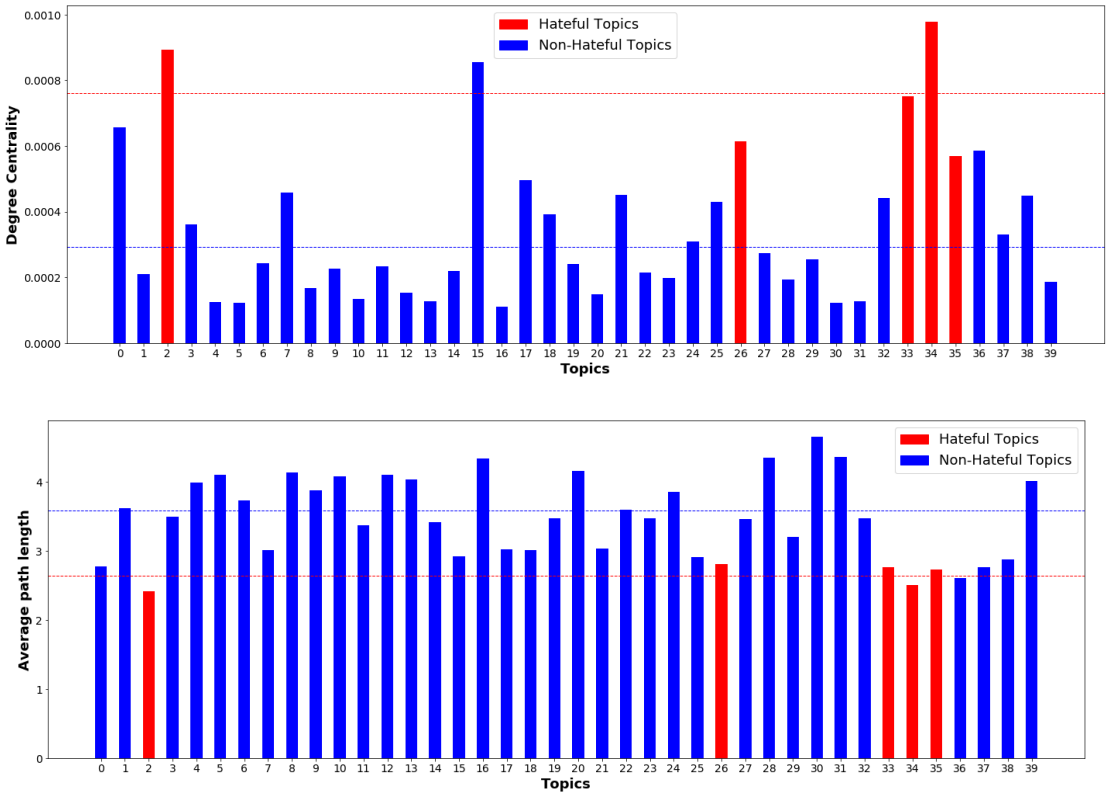

Figure 4: Distribution of Degree Centrality and Average Path Lengths.

mean average degree of the SCC of the hateful topics is 35.28 as compared 8.54 for non-hateful topics.

4. Conclusion and Future Work

We present a distinctive approach to study the dissemination of hate speech on Twitter. We combine a topic model with an ensemble based learning approach to detect hate speech in tweets. This allows us to capture the diverse nuances of hateful content. Our results indicate that propagation of hateful content can be effectively studied through a topic-based analysis of tweets. On analysing the roles of different types of users, we observe that there are a significantly higher number of propagators in topics that associate with hate. Further, examining the strongly connected components of the topical subgraphs provides insights into the community structure of hateful and non-hateful topics; observing that hateful topics have a denser and larger core, and hence we assume that information has a more lucid flow as compared to non hateful topics. This study, however, lacks in a temporal analysis of hate spread along multiple topics present. In the future, we plan to study the temporal spread of multiple forms of hate. Using the methodology and findings of this study, further analysis and experiments can reproach hateful content on OSNs rigorously.

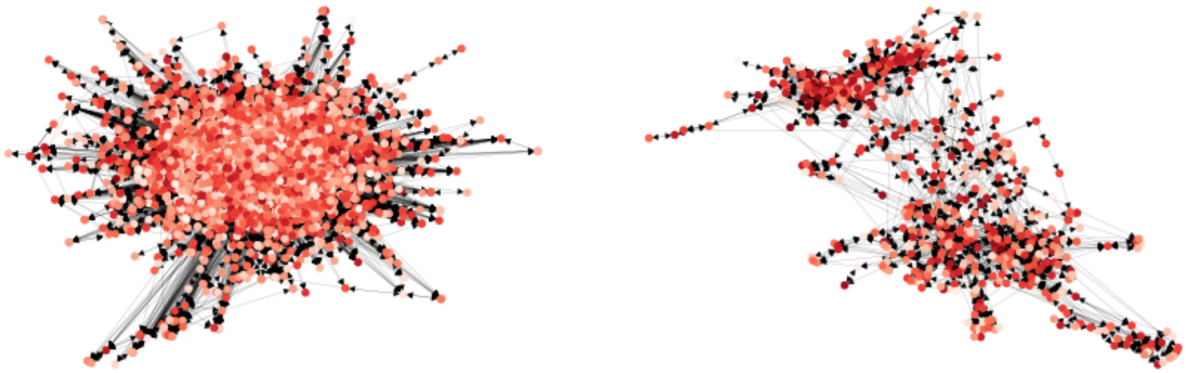

Figure 5: SCC Induced subgraphs for a representative hateful topic and a non-hateful topic. The average degree of users is 19.03 and 11.23 respectively.

References

[1] B. Ganesh, The ungovernability of digital hate culture, Journal of International Affairs 71 (2018) 30–49.

[2] B. Mathew, N. Kumar, P. Goyal, A. Mukherjee, et al., Analyzing the hate and counter speech accounts on twitter, arXiv preprint arXiv:1812.02712 (2018).

[3] B. Mathew, R. Dutt, P. Goyal, A. Mukherjee, Spread of hate speech in online social media, in: Proceedings of the 10th ACM Conference on Web Science, 2019, pp. 173–182.

[4] M. Ribeiro, P. Calais, Y. dos Santos, V. Almeida, W. Meira Jr, "like sheep among wolves": Characterizing hateful users on twitter, 2017.

[5] A. Arango, J. Pérez, B. Poblete, Hate speech detection is not as easy as you may think: A closer look at model validation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, 2019, p. 45–54.

[6] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, N. Asokan, All you need is "love": Evading hate speech detection, in: Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, AISec '18, 2018, p. 2–12.

[7] P. Ratadiya, D. Mishra, An attention ensemble based approach for multilabel profanity detection, in: 2019 International Conference on Data Mining Workshops (ICDMW), 2019, pp. 544–550. doi:10.1109/ICDMW.2019.00083.

[8] S. Wang, C. D. Manning, Baselines and bigrams: Simple, good sentiment and topic classification, in: Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2, Association for Computational Linguistics, 2012, pp. 90–94.