Empirical Assessment and Characterization of Homophily in Classes of Hate Speeches

Seema Nagar, [1] **Sameer Gupta,** [2] **C.S. Bahushruth,** [3] **Ferdous Ahmed Barbhuiya,** [1] **Kuntal Dey** [4]

[1] Indian Institute of Information Technology, Guwahati
[2] National Institute of Technology, Kurukshetra
[3] Manipal University, Jaipur
[4]Accenture Technology Labs, Bangalore
seema.nagar@iiitg.ac.in, sameer.lego@gmail.com, bahushruth.bahushruth@gmail.com, ferdous@iiitg.ac.in,
kuntal.dey@accenture.com

Abstract

In this paper, we investigate homophily in hate speech generation on social media platforms. Homophily plays a significant role in information diffusion, sustenance of online guilds, contagion in product adoption, the emergence of topics, and life-cycle on social networks. In the real world, features to utilize in similarity computation are well defined but not on social media platforms. We note that similarity among the users can be defined along with multiple aspects like: profile meta-data, the content generated and style of writing. These derived features are capable of capturing similarity along multiple dimensions, primarily semantic, lexical, syntactical, stylometric and topical. We leverage the important features for authorship attribution, word embeddings, latent and empath topics to compute lexical, syntactical, stylometric, semantic and topical features. We empirically demonstrate the presence of homophily on a dataset from Twitter along with the different aspects of similarity. Further, we investigate how homophily varies with different hateful types such as hate manifesting in topics of *gender*, *race*, *ethnicity*, *politics* and *nationalism*. Our results indicate higher homophily in users associating with topics of *racism* and *nationalism*.

Introduction

Social media platforms such as Twitter have enabled content generation by people in unprecedented ways, not imagined before. These platforms are now widely used to generate hate speech. Often times, the hate speech campaigns have incited real-life violence amongst people (Ribeiro et al. 2017; Mathew et al. 2018). There are many cases where countries have blamed social media platforms for inciting crimes in society. Facebook has been blamed for instigating anti-Muslim mob violence in Sri Lanka as well as for playing a leading role in the possible genocide of the Rohingya community in Myanmar. Therefore, studying multiple social aspects of hate speech such as diffusion, dissemination, and consumption is a critical problem.

(McPherson, Smith-Lovin, and Cook 2001) proposed homophily on social networks, using the assortative mixing hypothesis. Homophily on social networks is defined as "similarity breeds familiarity". It plays significant role in information diffusion on social networks (Aral, Muchnik, and Sundararajan 2009; De Choudhury et al. 2010; Halberstam and Knight 2016; Starbird and Palen 2012). The importance of homophily in information diffusion, motivated us to assess the presence of it in hate speech generation empirically. Works such as (Ribeiro et al. 2017; Mathew et al. 2019) study the positional aspect of hateful users in the social network. However, the literature has not explored homophily, a crucial aspect. We further investigate the strength of homophilic phenomenon for different types of hate such as, hate against gender, race, politics and ethnicity.

Two predominant factors are needed to assess homophily, familiarity, and similarity, which are naturally present on social media platforms. Familiarity captures the phenomenon of users becoming friends of (or, following) other users. The similarity is the phenomenon where a given user is similar to another user in the context of a given objective, such as generating hateful content or participating in the same topic. While familiarity on Twitter can be inferred using follower-followee or retweet social networks, similarity computation is not straight-forward. We believe the similarity between a pair of users should take multiple aspects of the content generated in addition to meta-data for the profiles. The multiple aspects, we explore in this paper are *semantic*, *syntactic*, *stylometric* and *topical*. We empirically investigate homophily along with the multiple aspects in hate speech generation.

We use word embeddings to compute *semantic* features for a user. The word embeddings are aggregated in a time-decaying manner to get a complete semantic representation of the user-generated content. We utilize the important features needed in authorship attribution (Bhargava, Mehndiratta, and Asawa 2013) and some other features designed by us to derive *syntactical* and *stylometric* features. Additionally, we also include readability (Kincaid 1975) related features. Lastly, we unearth the hidden thematic structure of a document along topics and categories in two ways, a) using latent topic modelling to construct a topic affinity vector and b) categories using Empath (Fast, Chen, and Bernstein 2016) to construct category score vector.

Hate speech constituents multiple types of hate. For example, hate against race, religion, ethnicity, gender, among

others. We believe that the different type of hate strength of homophily varies across different types of hate speech. We investigate the question, "Does the strength of homophily varies across the different types of hate?". We propose to use latent topic modelling to detect the types of hate present in a corpus.

In summary, we make the following contributions:

- We explore a slew of similarity features to capture the multiple aspects of user-generated content
- We experimentally investigate the usefulness of the various features, using homophily as the benchmark of comparison
- We do an in-depth analysis of variations in homophily strength across the different types of hate

Related Work

Many works have attempted to understand homophily on Twitter. (McPherson, Smith-Lovin, and Cook 2001) were the first to propose homophily in social networks. Subsequently, (De Choudhury et al. 2010) study the role of homophily in the diffusion of information on social networks. They build on the observation that homophily structures the ego-networks of individuals and impacts their communication behavior.

(Halberstam and Knight 2016) investigate the role of homophily in political information diffusion. (Aral, Muchnik, and Sundararajan 2009) show that homophily is also an important factor to explain contagion in product adoption on dynamic networks. (Ducheneaut et al. 2007) demonstrate that in the online gaming world, the sustenance of a gaming guild is driven by homophily. Thus, homophily has been very well studied in the literature and is very important to explain many social phenomena happening in the virtual world.

Many papers have jointly utilized similarity and familiarity for modeling solutions on Twitter and other online social networks. (Afrasiabi Rad and Benyoucef 2014) study communities formed over friendships on the Youtube social network. They observe that communities are formed from similar users on Youtube; however, they do not find large similarity values between friends in YouTube communities. Recently, topical homophily is proposed by (Dey et al. 2018), where they show the homophily is the driving factor in the emergence of topics and their life cycle. However, the existing literature does not at all address homophily in hate speech.

Central Idea

The proposed methodology has three main parts, a) features for similarity calculation, b) validating these features in homophily in hate speech and c) discovering types of hate in a corpus using latent topic modelling techniques. We argue that similarity calculation on social media platforms should capture multiple aspects of a user, instead just using direct textual similarity. These aspects are: user profile information, writing-related nuances such as stylometry, the content-generated itself and topics discussed, among others.

Algorithm 1 Computing Semantic Features for a User

Input: User u, Set of Posts $P = \{p_1, p_2, ..., p_M\}$ with timestamps
Output: Semantic Embedding $S(u)$ of the user

Compute time span of P as T
Divide T into time-windows $T = \{t_n, t_{n-1}, t_{n-2}, ..., t_1\}$ of size one week where n are the total number of weeks and t_1 is the most recent week
for each time window t in T **do**
    Compute $weight(t_k) = 1/k$
end for
for each post p in P **do**
    for each word w in p **do**
        Compute word embedding $E(w)$ using Glove
        Compute tweet embedding $E(p)$ as mean of word embeddings $E(w)$
    end for
    Find weight $W(p)$ for p using weight for the time windows it falls in
end for
Compute user semantic embedding $S(u)$
$S(u) = \sum_i E(p_i) * W(p)/|P|$

We show that homophily exists in hate speech generation on a dataset from Twitter along all the aspects utilized for similarity computation. Finally, we propose a topic modelling based approach to detect the different types of hate present in hate speech.

Features for Similarity Computation

We propose various features to capture the nuances of the content generated by a user on online social media platforms. The features are capable of capturing similarity along *semantic, syntactic, stylometric*, and *topical* dimensions.

Semantic Features We use word embeddings to represent user-generated content in a vector form. We get embedding for each post made by a user by taking the mean of the word embeddings and then aggregate the posts embeddings to get semantic embedding of the user. We use weighted mean pooling for performing aggregation. Aggregation methodology is motivated by (Rajadesingan, Zafarani, and Liu 2015), where the authors introduce the importance of time-decay in the content produced by a user. Time-decaying aggregation captures two crucial factors, a) Some users are more active than others and b) recent tweets are more important compare to the older ones. We split the tweets into time buckets of size one week. We assign a weight to the tweets in a bucket inversely proportional to its position in time. Formally, the time-decay based aggregation is described in1.

Syntactic Features We use important features proposed in (Rajadesingan, Zafarani, and Liu 2015) and some features designed by us to compute a syntactical feature vector for a user. These features include: number of capital words, question marks, exclamations, numbers, URLs, user mentions,

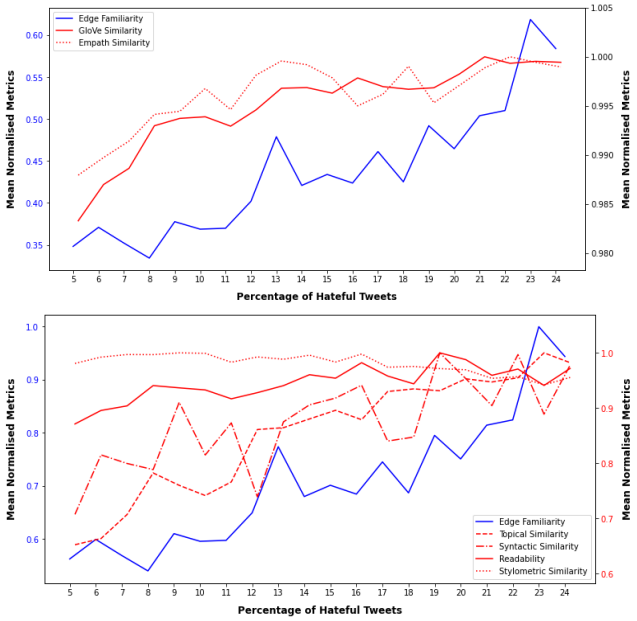

Figure 1: Variation in similarity and familiarity as hatefulness increases in community 1

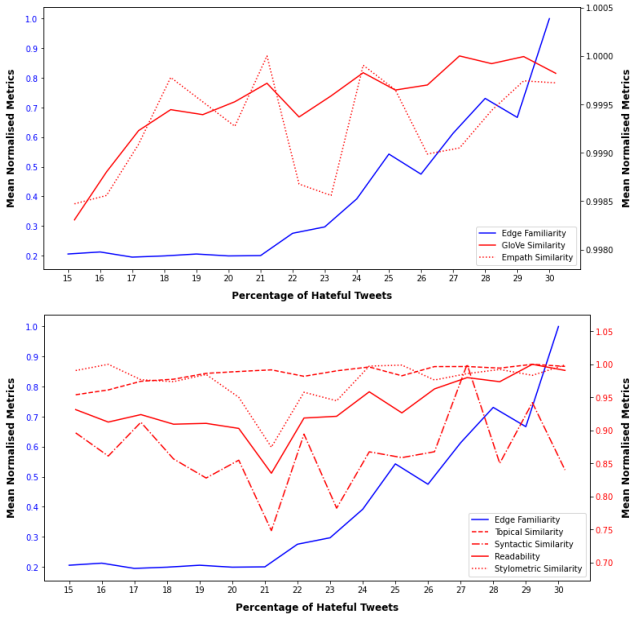

Figure 2: Variation in similarity and familiarity as hatefulness increases in community 2

hashtags, emojis, present in a tweet and then averaged over all the tweets posted by a user.

Stylometric and Readability Features

We use important features for authorship attribution from (Bhargava, Mehndiratta, and Asawa 2013) and some features from (Rajadesingan, Zafarani, and Liu 2015). Authorship attribution aims to detect the author of a piece of content produced, motivated us to use these features to capture the style of a user. These features include number of words per tweet, number of sentences per tweet, number of elongated words per tweet (e.g. hiiii), number of repeated words per tweet, word length distribution (vector of length 19 which has the frequency of words for that particular length), mean, median, the standard deviation of this distribution (Rajadesingan, Zafarani, and Liu 2015).

Additionally, we also compute the readability score for each user by using Flesch-Kincaid Reading Ease formula (Kincaid 1975). We create a document d for each user u by combining all the tweets as shown in Equation 1 and then perform readability computation.

Topical Features We compute topic features in two ways, a)perform topic modelling on the user-generated content. We employ latent topic modelling techniques, as described in the next section. We use the latent affinity to topics to construct a topic vector for each user and b) using the methodology proposed in (Fast, Chen, and Bernstein 2016), construct empath category scores vector.

Hateful Forms Detection using Topic Modelling

We use a latent topic detection technique called LDA (Blei, Ng, and Jordan 2003) to detect the latent topics present in a tweet. Due to tweets being short in length and large in number, scaling of LDA to detect topics where every tweet is treated as one document is very challenging. Therefore, we create one document per user by concatenating all his posts, which includes tweets, retweets, and quotes. Let a use u_i have made posts P, where $P = p_1, p_2, ..., p_N$. Then, the document d_i for user u_i is created by concatenating all the tweets in one document. Therefore, we have:

$$d_i = \cup_{(\forall p_j \in P)} p_j \tag{1}$$

Let $D = (\forall i \in 1..n) d_i$ be the corpus of documents. We further investigate D to detect latent topics present in it using LDA based techniques. We explore two variants of sampling for LDA, a) variational Bayes sampling method and b) Gibbs Sampling. Let the set of latent topics is T, where $T = t_1, t_2, ..., t_n$. The latent topic modelling produces a vector of topic affinity scores v_{d_i} for each document d_i. Let a_i is the affinity scores with respect to topic t_i, then we have topic affinity vector as follows:

$$T_{d_i} =< a_1, a_2, a_3, ..., a_T > \tag{2}$$

Experiments

Experiments Overview

The purpose of the experiments is to investigate the following research questions:

- *RQ1: Is homophily exhibited by the users generating hateful content and does it vary across the different types of similarity aspects?*
- *RQ2: Is homophily pronounced for particular hateful forms?*

Experiment Settings

Experimental dataset

We use the dataset provided by (Ribeiro et al. 2017). This dataset contains 200 most recent tweets of $100,386$ users, totaling to $19M$ tweets. It also contains a retweet induced graph of the users. It has $2,286,592$ directed edges. The dataset does not have labels for the tweet content. Therefore we manually annotate the tweets as hateful or not. Annotating $19M$ tweets of all the users is a costly and time-consuming process. Therefore, we pick only a sub-set of the users whose tweets we manually annotate. We run modularity optimization-based community detection using networkx[1] to pick a sub-set of users on the retweet network. The two communities picked have an equal number of edges around $1,60,000$ while the users are $7,679$ and $3,277$ respectively. These two communities have a sufficient number of users (from the perspective of the number of tweets to label) and edge density varies significantly between the two.

Parameter setting

We use existing familiarity metric of an edge existing or not, between a pair of users. We compute similarity in six different ways. *Semantic* features are constructed using glove word embeddings[2] while *syntactic and stylometric* are extracted based on (Bhargava, Mehndiratta, and Asawa 2013; Rajadesingan, Zafarani, and Liu 2015). We construct latent topical features by running Latent Dirichlet Allocation (LDA) using MALLET[3] on tweet corpus. The tweet corpus consists of tweets documents for all the users, wherein a tweets document is created for each user by concatenating all her posts. We set $\alpha = 5.0$ and $\beta = 0.01$. We use the library empath-client[4] to compute category score vector for each user using the tweets document.

We again use LDA to detect hateful topics. In this case, we only pick hateful tweets to create a tweets document for a user. We perform grid search where we vary alpha between 0.1 and 0.01 and the number of topics from 6 to 12. The number of iterations for each run is 500. We look at the coherence scores (Röder, Both, and Hinneburg 2015) and visualization of topics in terms of overlap using pyLDAvis[5]. We find that the best performing topic model, which has both a high coherence score and the least number of overlapping topics, is when $\alpha = 0.1$ and number of topics equal to 8. We observe that $\alpha = 0.01$ gives us a higher coherence score for the same number of topics as compared to $\alpha = 0.1$.

Experiments Results

To answer RQ1, we plot similarity, computed as cosine similarity, against familiarity for the six types of similarity metrics in Figures 1 and 2 for the community 1 and community 2 respectively. We vary the hatefulness of the users on the x-axis, where hatefulness of a user is defined as the percentage of hateful tweets. We see that as the hatefulness of the users increases, similarity values for all types of features also increases. We further observe that this pattern is enhanced in topic-based similarity. This is hinting that latent topics are capable of capturing the higher level semantics of the discussion happening on the social media platforms.

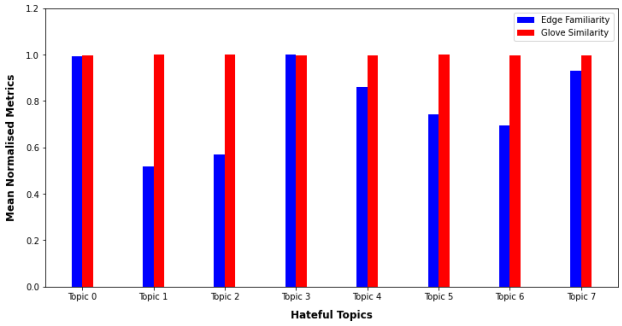

Figure 3: Variation in Homophily for the hate types in both the communities

To answer RQ2, we create a user base for each hate type(topic). We pick users whose affinity score is above a certain threshold. We also rank the different hashtags used by users by frequency. This is shown in Table 1. We observe that many users show higher values of association with specific topics $(0, 5, 7)$, as compared to the rest. Therefore, we decide to have a dynamic threshold for topic affinity. To compute these affinity thresholds, we select users in such a way that there are a reasonable number (at least 10% of total users) of representative users. For each topic, we plot the average familiarity, and average similarity in 3. The similarity and familiarity values are normalized by dividing by the maximum values, respectively. We observe that topics 3 and 7 exhibit stronger homophily, as compared to others for both the communities. These topics can be broadly categorized into hate manifesting *nationalism* and *racism*.

Table 1: Top Hashtags for the Hateful Topics

Topic	Hashtags
0	#maga, #trump, #realdonaldtrump, #trumptrain
1	#impeachtrump, #trump, #trumprussia, #jfkfiles
2	#bitch, #metoo, #harvey, #lockherup
3	#gobills, #pelicans, #mlscupplayoffs
4	#london, #fakenews, #cancer, #queen
5	#tormentedkashmir, #kashmirsuffering, #pakistan
6	#brexit, #crime, #terrorism, #illegal
7	#nigga, #bitch, #bitches, #somalia, #nigger

Conclusion

In this paper, we demonstrate homophily in hate speech on social media platforms. We also show that certain hate types exhibit stronger homophily in comparison to others. Unlike in the real world, features to compute similarity on social media platforms is not straightforward to define. Therefore, we propose a slew of features to capture similarity along

[1]https://networkx.org/

[2]https://github.com/stanfordnlp/GloVe

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

[4]https://github.com/Ejhfast/empath-client

[5]https://pypi.org/project/pyLDAvis/

with multiple aspects present on social media platforms. We demonstrate homophily in hate speech generation along with all these aspects. Further, we observe the variation of homophily in different classes of hate. We find that *racism*, and *xenophobia (nationalism)* shows stronger evidence of homophily among users.

References

Afrasiabi Rad, A.; and Benyoucef, M. 2014. Similarity and ties in social networks a study of the youtube social network. *Journal of Information Systems Applied Research* 7(4): 14.

Aral, S.; Muchnik, L.; and Sundararajan, A. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* 106(51): 21544–21549.

Bhargava, M.; Mehndiratta, P.; and Asawa, K. 2013. Stylometric Analysis for Authorship Attribution on Twitter. In *Big Data Analytics*, 37–47.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3: 993–1022.

De Choudhury, M.; Sundaram, H.; John, A.; Seligmann, D. D.; and Kelliher, A. 2010. "Birds of a Feather": Does User Homophily Impact Information Diffusion in Social Media? *arXiv preprint arXiv:1006.1702* .

Dey, K.; Shrivastava, R.; Kaushik, S.; and Garg, K. 2018. Assessing Topical Homophily on Twitter. In *International Conference on Complex Networks and their Applications*, 367–376. Springer.

Ducheneaut, N.; Yee, N.; Nickell, E.; and Moore, R. J. 2007. The life and death of online gaming communities: a look at guilds in world of warcraft. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 839–848.

Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4647–4657.

Halberstam, Y.; and Knight, B. 2016. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of public economics* 143: 73–88.

Kincaid, J. 1975. *Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Research Branch report. Chief of Naval Technical Training, Naval Air Station Memphis.

Mathew, B.; Dutt, R.; Goyal, P.; and Mukherjee, A. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, 173–182.

Mathew, B.; Kumar, N.; Goyal, P.; Mukherjee, A.; et al. 2018. Analyzing the hate and counter speech accounts on Twitter. *arXiv preprint arXiv:1812.02712* .

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27(1): 415–444.

Rajadesingan, A.; Zafarani, R.; and Liu, H. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *WSDM*, 97–106.

Ribeiro, M.; Calais, P.; dos Santos, Y.; Almeida, V.; and Meira Jr, W. 2017. "Like Sheep Among Wolves": Characterizing Hateful Users on Twitter. In *MIS2 Workshop at WSDM'2018*.

Röder, M.; Both, A.; and Hinneburg, A. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408.

Starbird, K.; and Palen, L. 2012. (How) will the revolution be retweeted? Information diffusion and the 2011 Egyptian uprising. In *Proceedings of the acm 2012 conference on computer supported cooperative work*, 7–16.