Empirical Assessment and Characterization of Homophily in Classes of Hate Speeches

AAAI'21 Workshop On Affective Content Analysis (AFFCON2021)

Background and Motivation

1. Homophily is defined as the tendency of like-minded (similar) people to connect/befriend (familiar)
2. Homophily structures a user's ego network on social networks
3. Homophily plays a significant role in information diffusion and dissemination
4. Homophily is a driver factor in product adoption, online guild formation, sustenance and community formation
5. But homophily has not been studied in generation of hate speech

Our Approach

The proposed approach has two main components:

1. **Defining features for similarity computation**:
   - We propose features which are capable of capturing similarity along semantic, syntactic, stylometric, and topical dimensions
   - Semantic Features are computed using the emebedding techniques
   - Syntatic Features captures Twitter related nuances such as number of capital words, question marks, exclamations, numbers, URLs, user mentions,
   - Stylometric features using authorship attribution
   - Topical Features constructed using two ways, a) topic modelling and b) empath scores

Let $\bar{s_1}$ and $\bar{s_2}$ represent similarity features for the users u_1 and u_2 respectively.

$$CosineSimiliarity(u1, u2) = \frac{(\bar{s_1} \cdot (\bar{s_2})}{||\bar{s_1}|| \cdot ||\bar{s_2}||} \tag{1}$$

Our Approach (cont’d)

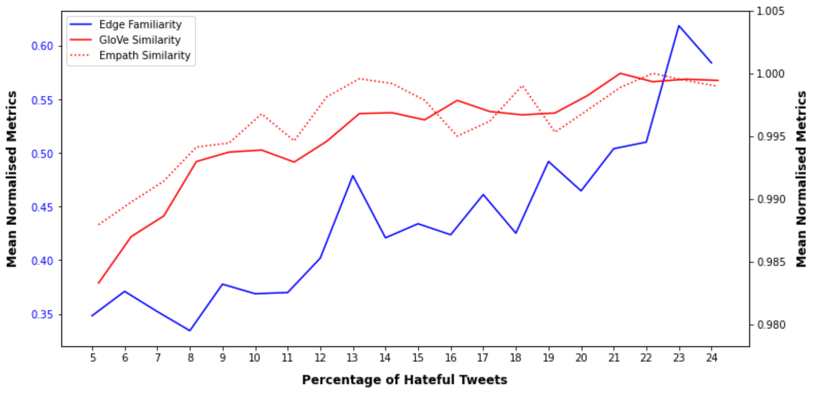
2. Detecting hateful forms on social media platforms:
    - We use latent topic modelling to detect multiple hateful forms present in hate speech
    - We hypothesize that individual hateful forms, differing in nature, might exhibit varied homophilic behaviours

Experiments

- Dataset
    - We use hate speech dataset provided by ””Like Sheep Among Wolves”: Characterizing Hateful Users on Twitter”
    - It has 200 recent tweets of 100,386 users along with retweet induced graph
    - We pick a sub-set of the users whose tweets we manually annotate
    - Use modularity optimization to detect community structure and pick two communities
    - Edge density varies significantly across the two communities
- 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?*

Experimental Results

- To answer RQ1 we plot similarity against familiarity for the six types of similarity metrics for both the communities
- As the hatefulness increases, homophily also increases.
- This pattern is enhanced in topic-based similarity.


Edge Familiarity
GloVe Similarity
Empath Similarity
Mean Normalised Metrics
0.35
0.40
0.45
0.50
0.55
0.60
Mean Normalised Metrics
0.980
0.985
0.990
0.995
1.000
1.005
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
Percentage of Hateful Tweets

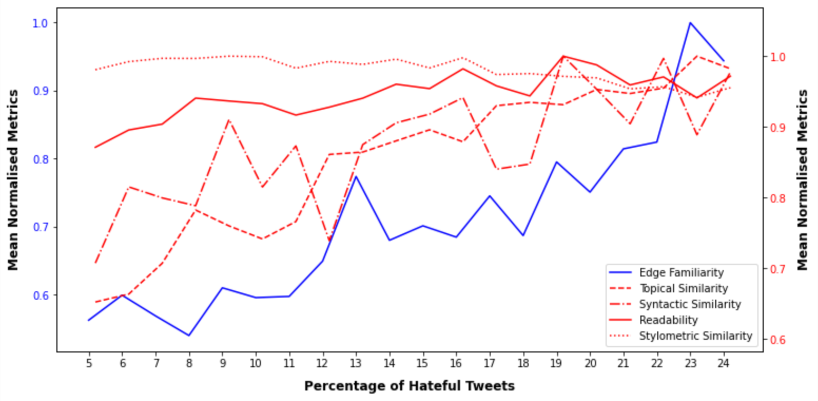
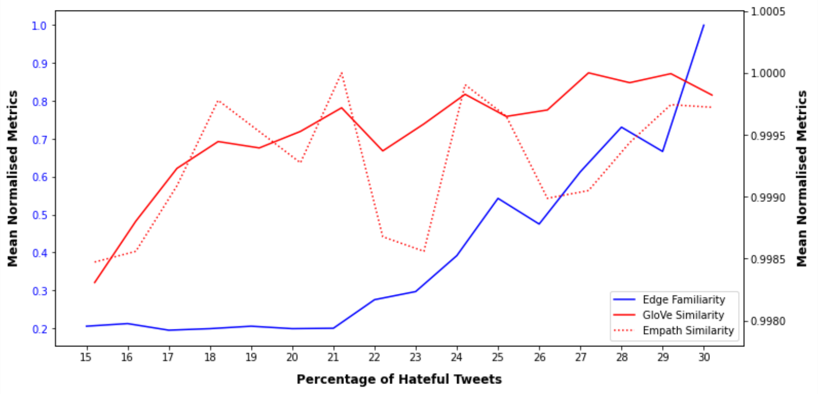


Figure 1: Variation in similarity and familiarity as hatefulness increases in community 1



Mean Normalised Metrics
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
Percentage of Hateful Tweets
1.0005
1.0000
0.9995
0.9990
0.9985
0.9980
Mean Normalised Metrics
Edge Familiarity
GloVe Similarity
Empath Similarity

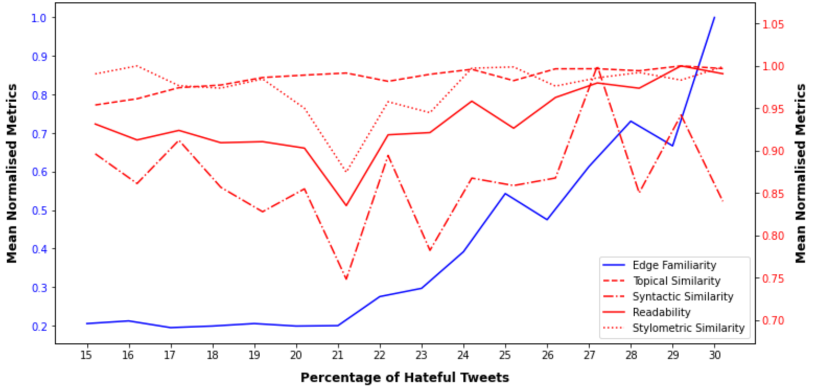


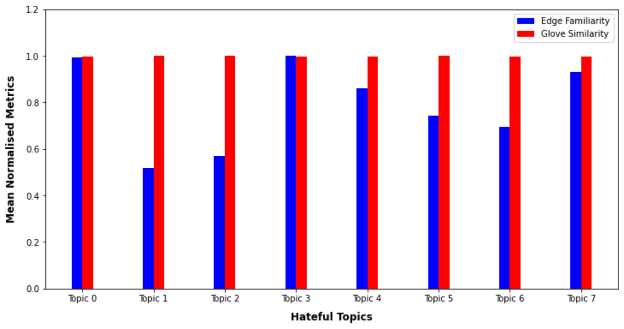
Figure 2: Variation in similarity and familiarity as hatefulness increases in community 2

- 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.
- For each topic, we plot the average familiarity, and average similarity

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



1.2
1.0
0.8
0.6
0.4
0.2
0.0
Mean Normalised Metrics
Edge Familiarity
Glove Similarity
Topic 0
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Hateful Topics

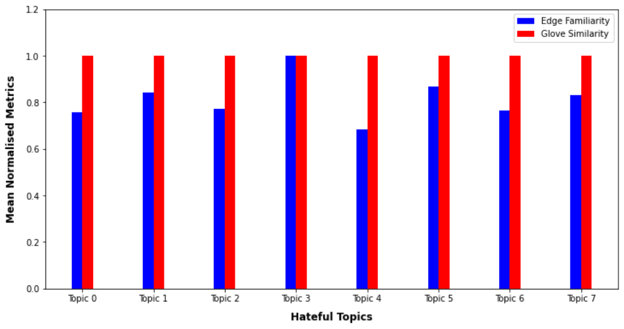


Figure 3: Variation in Homophily for the hate types in both the communities

Summary

1. We propose a novel metrics to compute similarity
2. We show homophily in hate speech on a dataset from Twitter.
3. We empirically demonstrate the effectiveness of the newly proposed metrics in establish similarity against the existing metrics, using homophily as the benchmark of comparison.
4. We do a deep dive analysis of variations of homophily in different forms of hate.