Haters Gonna Hate

Check-in Report 3

Aviral Chawla[1], Daniel Orem[2], Jay Hwasung Jung[3], and Shunsuke Miyazato[4]

[1]Complex Systems and Data Science (CSDS) M.S., University of Vermont
[2]Chemistry (CHEM) B.S., University of Vermont
[3]Computer Science (CS) B.S., University of Vermont
[4]Data Science (DS) B.S., University of Vermont

Nov 27, 2022

Abstract

Social media is widely used across continents, generations, and social groups. Due to the accessibility and anonymity of social media, hate speech, an abusive or threatening statement showing prejudice and hate, has become a new social phenomenon and has gotten public attention. In previous research, there has been an effort to understand the behavior of such hateful speech using Natural Language Processing and Networks, and social media platforms have hate-speech detection models based on such research. Although research provides an effective method to detect hateful speech, research understanding its behavior still needs to be completed. This project aims to provide the framework for social media platforms to conduct an early intervention in hate speech using network behavior of content instead of relying on language models for classification.

1 Activities

1.1 Planned Activities

For week 03 and week 04, our goal was to polish classification process get datasets, and assess model accuracy and start building a network.

1.2 Accomplished Activities

For week 03 and week 04, we tried to clean our datasets, and save it as a .db format, so we can retrieve datasets anytime we want. Also, last week, we did some machine learning model testings. Moreover, we found the API called Google Perspective which evaluates the score of texts in requested categories such as Toxisity, etc...

Below is the scores listed using Google Perspective API. These scores indicate probabilities that the given texts are classified as a category such as Toxicity, Insult, or Treat.

```
I will kill you
TOXICITY:  0.93383175
INSULT  :  0.28791866
THREAT  :  0.92211
```

Figure 1: Google Perspective API scores of three categories

We are expecting to compare this model with machine learning models we built and compare the effectiveness of them.

2 Challenges

2.1 Open Challenges and Questions

Due to the time constraints during the Thanksgiving, it was hard to communicate with teammates. Also, getting an "increasing quota" for Google Perspective API takes long. Originally, this API can request 60 times/minutes, but it will take us overnight to analyze all datasets we have; if there is an update in our model, we cannot help re-running the evaluation and classification.

2.2 Major Changes

Networks will be built next week.

3 Timeline and Roles

3.1 Timeline

Time	Task
Week 01	Literature review finish project proposal
Week 02	Extract data and work on cleaning and organizing it
Week 03	Polish classification process and get a final model-tagged data to work with
Week 04	Assess model accuracy and start building the network
Week 05	Network analysis and visualization
Week 06	Compile results and prepare final report

3.2 Roles

Name	Tasks
Aviral Chawla	Build Hate speech Network based on the classification of our models.
Daniel Orem	Analyze classified data in-depth
Jay Hwasung Jung	Build Hate Speech network and housekeep models.
Shunsuke Miyazato	Analyze classified data in-depth