Project Proposal

Arjun Mahadkar
B00976832
amahadk1@binghamton.edu

Deepang Raval*
B00924269
draval1@binghamton.edu

Rahul Verma
B00892091
rverma4@binghamton.edu

Sudeep Rawat
B00852066
srawat1@binghamton.edu

Yuraj Vartak
B00866245
yvartak1@binghamton.edu

1 INTRODUCTION

Various social media platforms currently are facing issues related to toxicity and misinformation. We plan to conduct experiments on tweets from Twitter, comments on posts of a subreddit and comments on videos posted on a YouTube channel to detect and understand the patterns associated with such behavior.

We plan to answer three research questions for each of the social media platform we have collected the data from.

1.1 Twitter Tweets Data

(1) **RQ1:** What percentage of tweets that are made related to an athlete, a sports team, a sport etc. are toxic?
(2) **RQ2:** For each named entity recognized from the tweets, what is the dominant sentiment towards it?
(3) **RQ3:** What are the factors that led to toxicity or negative sentiment?

1.2 Subreddit Comments Data

(1) **RQ1:** What percentage of comments that are made related to the U.S. politics are toxic?
(2) **RQ2:** For each named entity recognized from the comments, what is the dominant sentiment towards it?
(3) **RQ3:** What are the factors that led to toxicity or negative sentiment?

1.3 YouTube Comments Data

(1) **RQ1:** What percentage of comments that are made on a YouTube video are toxic?
(2) **RQ2:** What is the pattern of comments that are spam?
(3) **RQ3:** What is the structure of comments that are posted with the motive of scamming others?

2 PROPOSED METHODOLOGY

(1) To classify the tweets and comments into various classes like toxic, severe toxicity, insult etc. we would use the Perspective API[1]

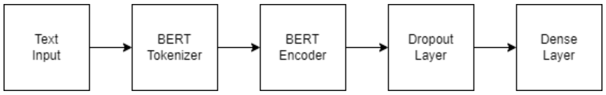

Figure 1: Methodology

(2) For automatic named entity recognition we would use the pretrained BERT model for token classification[2]
(3) We would also design a binary classifier for classifying the sentiments as positive or negative. The general architecture of this classifier is shown in figure 1
(4) Another classifier would be built to distinguish whether the input text is spam or ham. This classifier would help in distinguishing and visualizing spam text from our data. The architecture would be same as shown in figure 1
(5) The final classifier would be used to differentiate the comments that are meant to scam from those that are not. We would be using 80% of the scam comments from the data collected to train the model and 20% to test the accuracy of the model. This classifier also follows the same architecture as shown in figure 1
(6) All the outputs from step 1 to 5 would be visualized using Matplotlib[3] plots

3 DATA COLLECTION VALIDATION

We display the number of documents that are collected by the data collection system for all the social media platforms till 11/06/2022 18:30:00 UTC

(1) Twitter tweets data: 466900
(2) Subreddit comments data: 362688
(3) YouTube comments data: 447239

As large language models like BERT require significantly less amount of data to train on, we are optimistic that the current amount of data is suffice for our visualizations and as time proceeds we would accumulate at least double the amount of current data and hence won't require to collect any additional data.

[1] https://perspectiveapi.com/

* https://scholar.google.com/citations?user=MfAFu6QAAAAJ.

[2] https://huggingface.co/transformers/v3.0.2/model_doc/auto.html#tfautomodelfortokenclassification

[3] https://matplotlib.org/