Project Report

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

ABSTRACT

Social media platforms are proven to be a good source of data for scientific analysis and for training various machine learning models. We present exploratory data analysis of tweets from Twitter, comments on posts of a subreddit, and comments on videos posted on a YouTube channel. One of the issues social media platforms currently are facing is related to toxicity and misinformation. We show tables and plots to get a basic understanding of the patterns associated with such behavior. We also show the influence of various events on the collection of data. We present research questions based on the understanding of the data for future work on designing classifiers that can easily detect toxicity and misinformation.

1 INTRODUCTION

Exploratory data analysis refers to performing initial investigations on data to discover patterns and check assumptions with the help of summary statistics and graphical representations.

The data that we are exploring is being continuously collected by data extraction systems designed for each of the three social media platforms.

We use pretrained BERT [2] language model fine-tuned by Detoxify [4] to predict toxicity score of the given text. The categories it is divided into are toxic, severe_toxic, obscene, threat, insult, identity_hate.

BERT tokenizer is use-full in tokenizing the text provided to it. By using the pretrained BERT model for automatic token classification we generate named entities labels for the location, organization or person name present in the text.

We generate word clouds for named entities and spam/scam comments which is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it's mentioned within a given text and the more important it is.

Data is exported from MongoDB in a csv file and processed using pandas library and datetime is used to handle the date and time column present in the data. Seaborn library helps in generating plots to visualize the processed data.

2 LITERATURE SURVEY

Badjatiya etal. [1] conducted experiments on benchmark twitter dataset of 16K annotated tweets and showed that deep learning based methods for hate speech detection are better than char/word n-gram methods.

Zhao etal. [6] made use of BERT, RoBERTa, and XLM for toxic comment classification and showed that BERT and RoBERTa generally outperform XLM on toxic comment classification.

Gong etal. [3] showed how BERT can be fine-tuned to be used for Named Entity Recognition(NER) and presented some ways to improve its accuracy.

Li etal. [5] proposed dice loss in replacement of the standard cross-entropy objective for data-imbalanced NLP tasks like NER and showed improved accuracy for BERT based models using it.

3 EXPLORATORY DATA ANALYSIS

3.1 Twitter Tweets

(1) Table 1 shows the examples of the tweets that had high toxicity scores. We can see that users try different ways so that the toxic tweet is not directly detected. Also the sport event on that day lead the user to post a toxic tweet.

(2) Figure 1 shows the number of tweets the data extractor was able to fetch for our desired contexts. Most of the tweets are related to a sports event. Tweets starting from 11/01/2022 to 11/14/2022 are used for the count.

(3) Figure 2 shows how the events of the day influence the tweets context. Since on 11/01/2022 WWE aired it had more tweets while on 11/02/2022 there was an ind vs ban cricket match due to which it had more tweets.

(4) Figure 3 shows the named entities that were recognized from the tweets on 11/01/2022 and 11/02/2022. This shows that tweets had the term nfl and nba but they were not in the top 5 hashtags hence we got to know something new about our data.

(5) Figure 4 shows the number of tweets that came in over the sample stream starting 11/01/2022 through 11/14/2022.

Table 1: Toxic Tweets

Tweet	toxic	threat	obscene	identity_hate	insult	severe_toxic
@ChelseaFC #BlackHistoryMonth All the mf s l a v es should be die in this club	0.905	0.349	0.244	0.243	0.185	0.045
That #nigger #justinfields needs to go down. Fuck him and those faggots #lgbqt	0.999	0.383	0.988	0.901	0.976	0.750
Adam Silver, how's this for an apology? Fuck off, jew. @tmz https://t.co/SGutlUBQHM	0.996	0.015	0.975	0.673	0.888	0.326
@babarazam258 and @iMRizwanPak just trapped pakistan to lose by wasting time. Selfish self centerer morons who play for themselves not for a pakistan	0.828	0.002	0.204	0.156	0.520	0.009

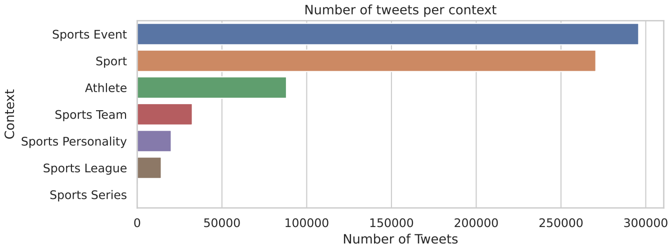

Figure 1: Context vs. Number of Tweets

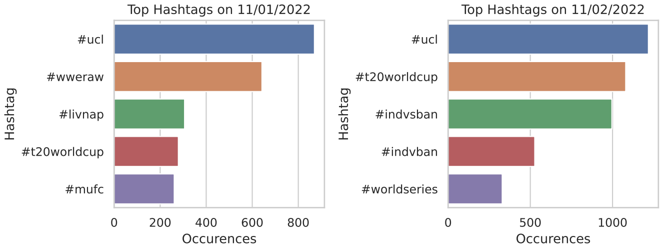

Figure 2: Hashtags vs. Occurrence in Tweets

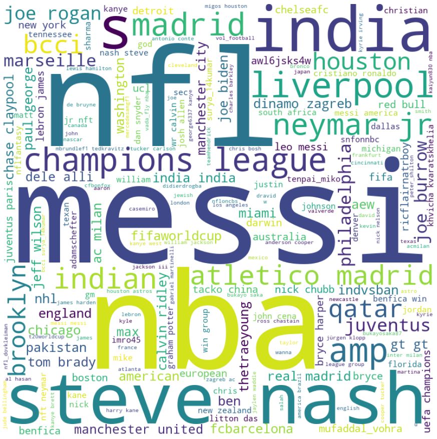

Figure 3: Word Cloud of Named Entities from Tweets

3.2 Reddit Comments

(1) Table 2 shows the examples of the comments posted on a post on the subreddit that had high toxicity scores.

(2) Figure 5 shows the number of comments the data extractor was able to fetch for our desired subreddits. Comments starting from 11/01/2022 to 11/14/2022 are used for the count.

(3) Figure 6 shows the named entities that were recognized from the comments posted on 11/01/2022 and 11/02/2022. Republican appeared to be the most common word.

(4) Figure 7 shows the number of submissions per day on the r/politics subreddit from 11/04/2022 to 11/14/2022.

Table 2: Toxic Reddit Comments

Comment	toxic	threat	obscene	identity_hate	insult	severe_toxic
This scotus is ILLIEGILIMATE! FUCK them, and fuck anyone who supports them. The government should IGNORE their rulings (read: BULLSHIT)	0.998	0.005	0.990	0.098	0.958	0.370
I admire your optimism but I wouldn't be so sure about that. The left has been letting Republicans fuck them up the ass for years now. I don't see why this would be any different.	0.940	0.011	0.855	0.187	0.447	0.053
The GOP and it's followers are a soulless cult. Save the babies, but hurt and kill any opponents. They were going to hang Mike Pence. We heard nothing from the GOP because they are brainwashed and cowards.	0.949	0.148	0.115	0.246	0.579	0.027

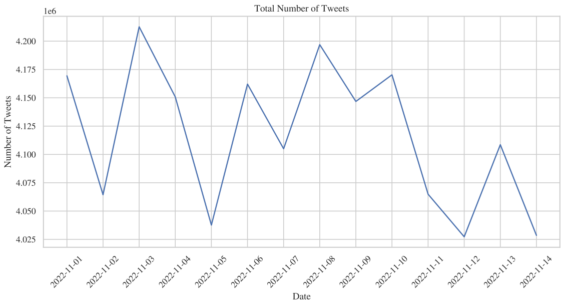

Figure 4: Total Number of Tweets

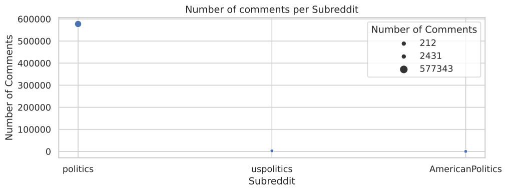

Figure 5: Subreddit vs. Number of Comments

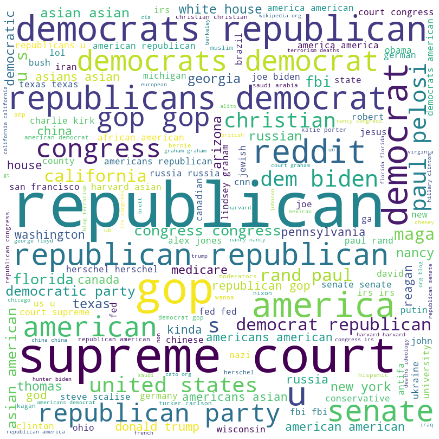

Figure 6: Word Cloud of Named Entities from Subreddit posts Comments

3.3 YouTube Comments

(1) Table 3 shows the examples of the comments posted on a video that had high toxicity scores.

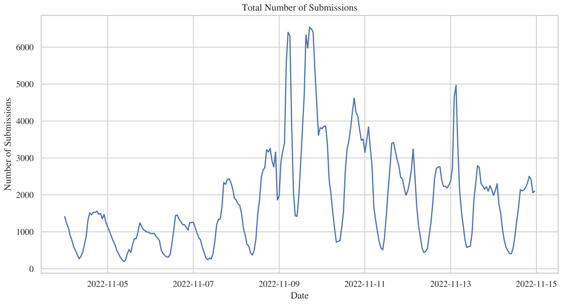

Figure 7: Total Number of r/politics Submissions

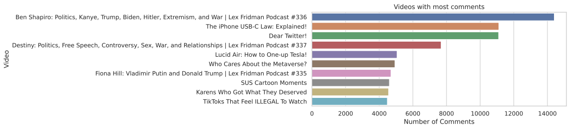

Figure 8: Videos with most Comments

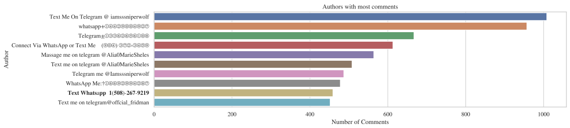

Figure 9: YouTube Authors with most Comments

(2) Figure 8 shows the videos with most comments starting from 11/01/2022 to 11/14/2022. It proves that videos on hot topic or controversial personality receives more comments.
(3) Figure 9 shows that most of the comments on videos are made by authors with the intention of spam/scam.
(4) Figure 10 shows the most frequent words used in the comments that are intended for spamming/scamming.

3.4 Common Plots

(1) Figure 11 shows that each social media platform follows the same pattern through a day when it receives more comments at some hour and less at some.
(2) Figure 12 shows the influence of an event on number of submissions. Due to US Elections on 11/08/2022 there was a sudden spike in number of comments in us-politics related subreddits and later in the plot Fifa World cup influenced the increase in number of tweets.

4 DISCUSSION AND CONCLUSION

We saw that there is toxicity present in each of the social media platform regardless of the context of it. We showed that BERT based toxic comments classifier can fairly score between the various toxic categories. There are specific trends in submissions that are influenced by an event. BERT based NER approach was able to generate entities based on person name, organization and location.

Current BERT based classifiers have a limitation of being able to process up to 512 tokens hence sequences with more than 512 tokens couldn't be processed which can lead to incomplete information while understanding the data. Also these classifiers are not much scalable and can result in a

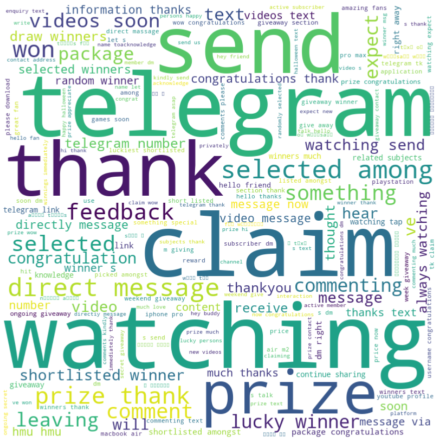

Figure 10: Word Cloud of most frequent words in YouTube comments related Spam/Scam

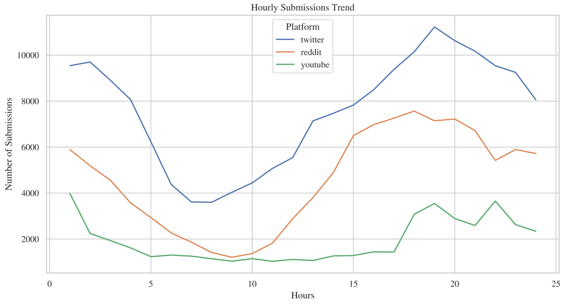

Figure 11: Hourly Submissions Trend across all the platforms

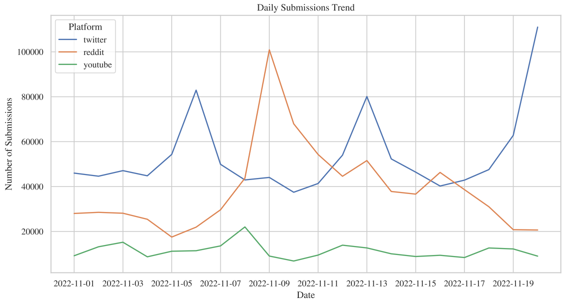

Figure 12: Daily Submissions Trend across all the platforms

Table 3: Toxic YouTube Comments

Comment	toxic	threat	obscene	identity_hate	insult	severe_toxic
Hes a grown adult, why the fuck would he want to watch a kid sitting in his room screaming?	0.996	0.004	0.978	0.008	0.705	0.151
who the fuck watches this	0.992	0.005	0.979	0.003	0.268	0.182
wtf are they deleting my comments i will post it everywhere if i die its ur fault	0.941	0.075	0.427	0.011	0.046	0.022
Sidemen would 100% be better if they didn't hang on to KSI's reputation, this is why Betasquad is better. KSI is boring and peak as fuck.	0.958	0.002	0.897	0.007	0.384	0.029

higher processing time on a low-end GPU or CPU that makes it hard to use when processing large chunks of data.

These issues could be resolved by using the various other language model architectures that support more number of tokens and could be scalable easily. We would use such an architecture for future analysis of the data that has proven to be more accurate than the vanilla BERT architecture.

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

4.1 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?

4.2 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?

4.3 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?

REFERENCES

[1] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 759–760. https://doi.org/10.1145/3041021.3054223

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[3] Yuan Gong, Lu Mao, and Changliang Li. 2021. Few-shot Learning for Named Entity Recognition Based on BERT and Two-level Model Fusion. *Data Intelligence* 3, 4 (10 2021), 568–577. https://doi.org/10.1162/dint_a_00102 arXiv:https://direct.mit.edu/dint/article-pdf/3/4/568/1968571/dint_a_00102.pdf

[4] Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

[5] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2019. Dice loss for data-imbalanced NLP tasks. *arXiv preprint arXiv:1911.02855* (2019).

[6] Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2021. A Comparative Study of Using Pre-Trained Language Models for Toxic Comment Classification. In *Companion Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 500–507. https://doi.org/10.1145/3442442.3452313