Dataset for Hate Speech and Profanity Detection with Sub-Categorization

Barha Meherun Pritha, Tabassum Alam, Samin Islam
Md Rayhan Kabir, Nazmus Sakeef, Shahriar Rahman Rana

Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Abstract. Spreading hatred via social media platforms has been skyrocketing in recent years, as it's the easiest way to communicate all across the globe. The Influence of these online platforms is being used in ways that shape society. Hence, it has become an alarming concern to look over the text data online to maintain a safe space for communication. The dataset that has been provided with this article is extracted from Twitter, which are tacked down from the tweets of the general public as well as the famous figures. The purpose of this dataset collection is to detect the hate speech and profanity that's present at the tip of our fingertips with the help of Machine Learning techniques. This article has also given an elaborative categorization of the labeled hate and profanity tweets from different perspectives of discrimination. The regular tweets have also been classified into some categories.

1 Introduction

Hate speech and abusive language has been ever present and persistent to humankind for the longest time. Due to the access of the internet at a large scale by individuals, a number of online platforms including Twitter, Facebook, Instagram and YouTube are being used daily. With the aid of such platforms, people can leave hateful or offensive statements through posts, comments, chats, etc. from the safety of their homes. It has become a challenge for networking sites to establish a platform where no hurtful, contempt or profane language can take place. Online media are establishing and updating their policies regarding what kind of content or comments can be shared, but even then, it seems that not all hate speech can be filtered out. Even though human beings can identify hate speech through words, tones and sentence structure, it is a bit difficult for artificial intelligence to evaluate if anybody encourages hatred or just explains what has happened to them. With the help of Machine Learning methods researchers aim to overcome this drawback recently, and this article has a similar motive as well.

The given dataset is a combination of extracted English tweets by an API and labeled as hate speech, profanity, regular speech along with both hate speech and profanity. The theme of offensive speech is becoming high in recent times yet further progress has been made to discover how to better recognize and discriminate between hate speech and other normal speech. We can trace profanity or receptive conversation via the use of clear and unambiguous keywords but not necessarily hate speech. On that account the four labeled groups are also categorized into specific sub-groups of harassment and friendly tweets.

This approach will lead to more accurate identification of hate speech from the tweets. The dataset can be useful to train and test computational models and techniques of automatic classification.

2 Dataset Preparation

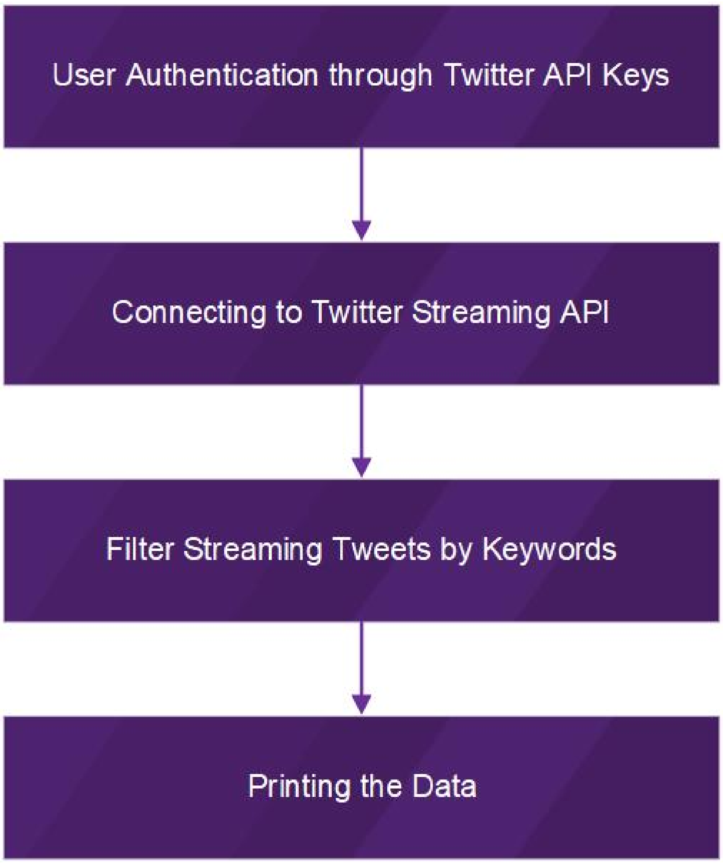

User Authentication through Twitter API Keys
Connecting to Twitter Streaming API
Filter Streaming Tweets by Keywords
Printing the Data

Fig. 1. The Process of our Data Collection

We have collected the data using the Python library tweepy. We have used the OAuthHandler submodule to get authentication from Twitter. Later, we used the Stream submodule to filter the tweets using keywords and then store them in a file. The process of data collection from Twitter is shown in **Fig. 1**.

The keywords we used to filter the tweets are: ‘food’, ‘election’, ‘media’, ‘competition’, ‘vlog’, ‘travel’, ‘USA’, ‘US’, ‘economy’, ‘politics’, ‘programming’, ‘social’, ‘climate’, ‘game’, ‘tournament’, ‘movie’, ‘culture’, ‘torture’, ‘trump’, ‘biden’, ‘show’, ‘finance’, ‘stories’, ‘marketing’, ‘media’, ‘twitter’, ‘facebook’, ‘research’, ‘disaster’, ‘weather’, ‘life’, ‘motivation’, ‘fitness’, ‘science’, ‘goals’, ‘technology’, ‘festival’, ‘concert’, ‘song’, ‘review’, ‘hate’, ‘love’, ‘romance’, ‘beautiful’, ‘scenario’, ‘place’, ‘football’, ‘cricket’, ‘computer’, ‘religion’, ‘feminism’, ‘job’, ‘study’, ‘worst’, ‘shut’, ‘racism’, ‘kill’, ‘slang’, ‘gun’,

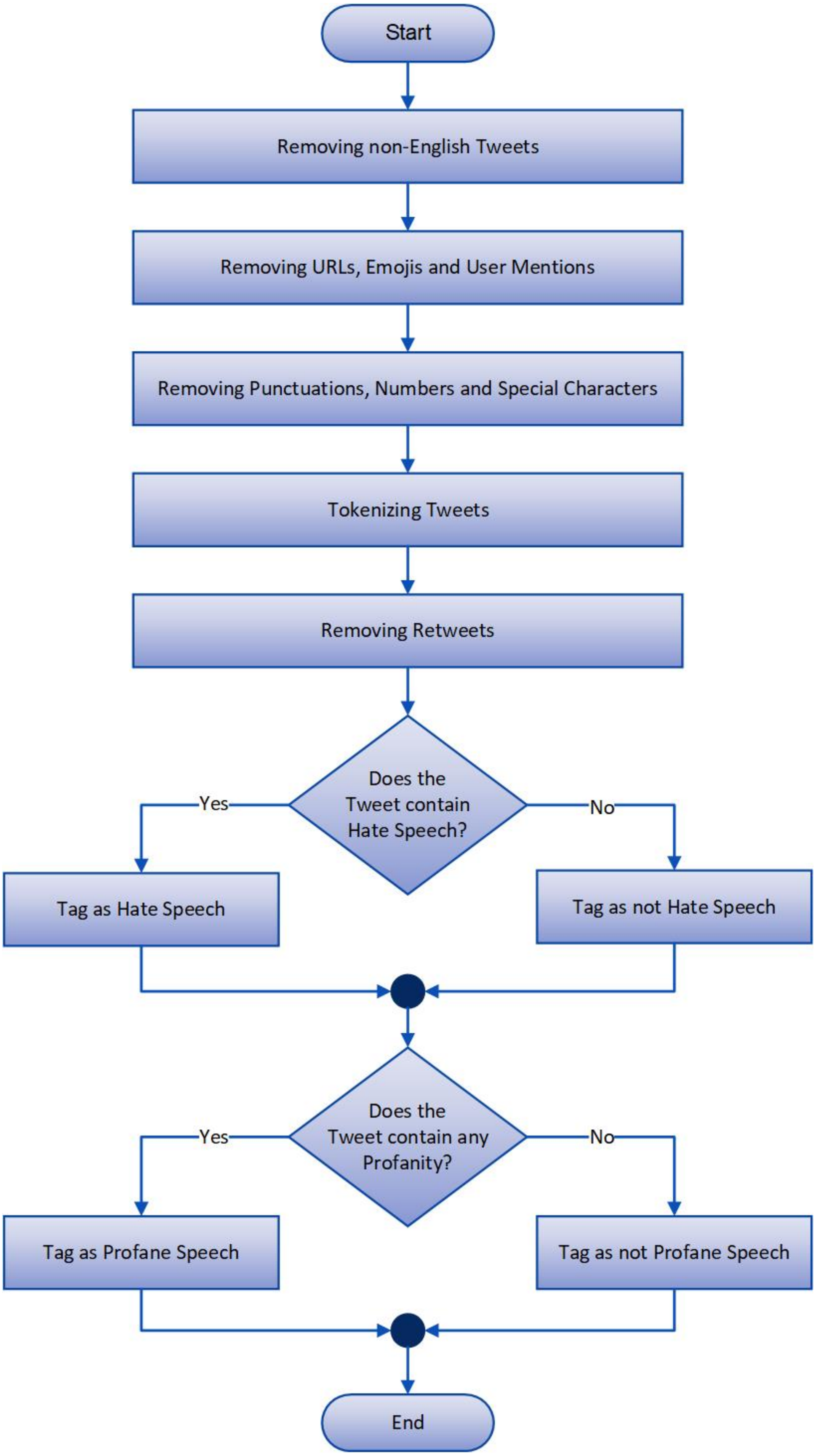

Start
Removing non-English Tweets
Removing URLs, Emojis and User Mentions
Removing Punctuations, Numbers and Special Characters
Tokenizing Tweets
Removing Retweets
Does the Tweet contain Hate Speech?
Yes
No
Tag as Hate Speech
Tag as not Hate Speech
Does the Tweet contain any Profanity?
Yes
No
Tag as Profane Speech
Tag as not Profane Speech
End

Fig. 2. Dataset Cleaning and Tagging

‘murder’, ‘suicide’, ‘racist’, ‘shit’.

After we have collected our data, as shown in **Fig. 2**, we have cleaned the data by removing non-English tweets, urls, emojis, mentions, punctuation, numbers and other special characters. Then we have removed all the retweets and have tokenized each word of the tweets. Then we have tagged them separately on the basis of - i) if they contain hate speech or not, and ii) if they contain any profanity or not.

3 Data Description

The **Table 1** describes how the datasets are divided into four categories; Combined, Hate Speech, Profanity and Regular. Combined, Hate Speech and Profanity are further divided into their types of discrimination; Bully, Racism, Belief, Politics, Sexual, Self Hatred and Criticism. Regular Speech is divided into Thoughts, Grateful, Inspiring, Praising and Wishes.

Categories	**Labels**
Combined	Bully, Racism, Belief, Politics, Sexual, Self Hatred and Criticism
Hate Speech	Bully, Racism, Belief, Politics, Sexual, Self Hatred and Criticism
Profanity	Bully, Racism, Belief, Politics, Sexual, Self Hatred and Criticism
Regular	Thoughts, Grateful, Inspiring, Praising and Wishes

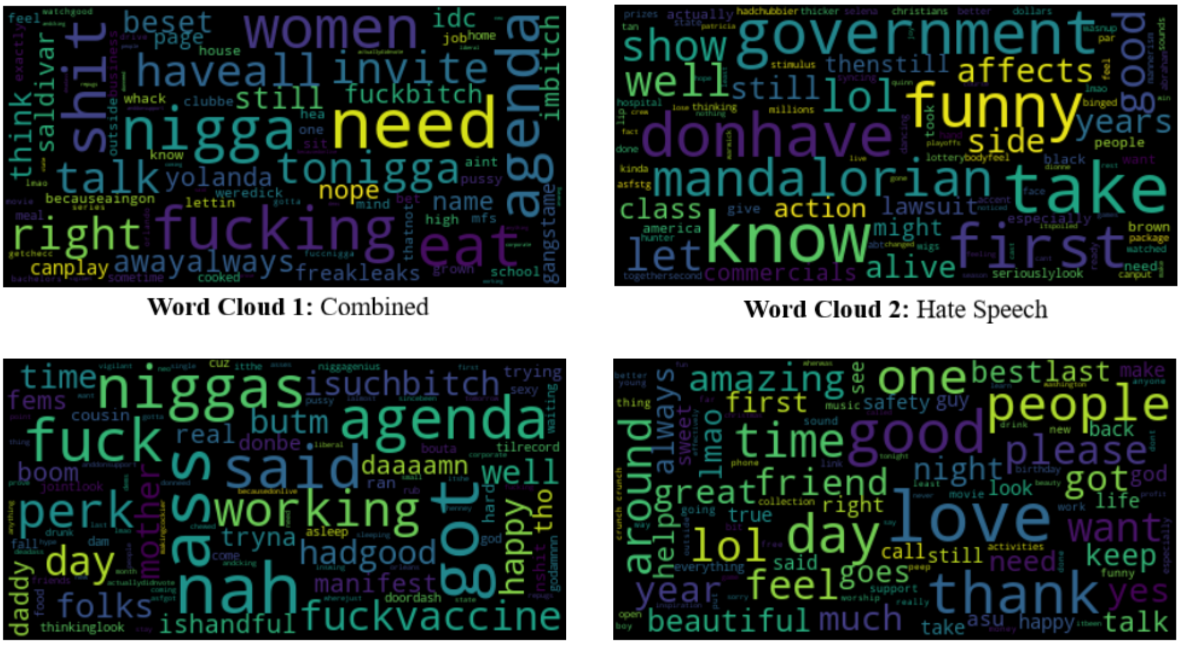
Table 1. Main Categorizations and their Types

4 Data Analysis

Fig. 3 illustrates the different Word Clouds generated from Tweets of different categories. They were coded with the help of python libraries in kaggle.

Fig. 4 illustrates the percentage of hate speech, profanity, combination of hate speech and profanity and friendly tweets based on the total number of data collected. From the pie chart it is obvious that above 50% of the total tweets contain hate speech and profanity, with just 2% difference between the two. Surprisingly, the percentage of tweets containing both hate speech and profanity are also above 15% which results in the total number of hateful tweets as 62%. Only 38% of the tweets refer to the friendly ones which contain no hateful contents.

In **Fig. 5**, the bar chart depicts the proportion of tweets that contain hate speech depending on the label manually assorted to determine if a tweet depicts “self hatred”, “beliefs”, “sexual”, “racist”, “politics”, “bully” or “criticism”. Almost half of the total tweets i.e. 42.20% of them portrays criticism. Negativity towards politics and bullying also consists of 49.54% of the total tweets. A few proportion of around 8% of the tweets depicts self-hatred, hatred towards beliefs, sexual and racism.



Word Cloud 1: Combined

Word Cloud 2: Hate Speech

Word Cloud 3: Profanity

Word Cloud 4: Regular

Fig. 3. Word Clouds of different Categories

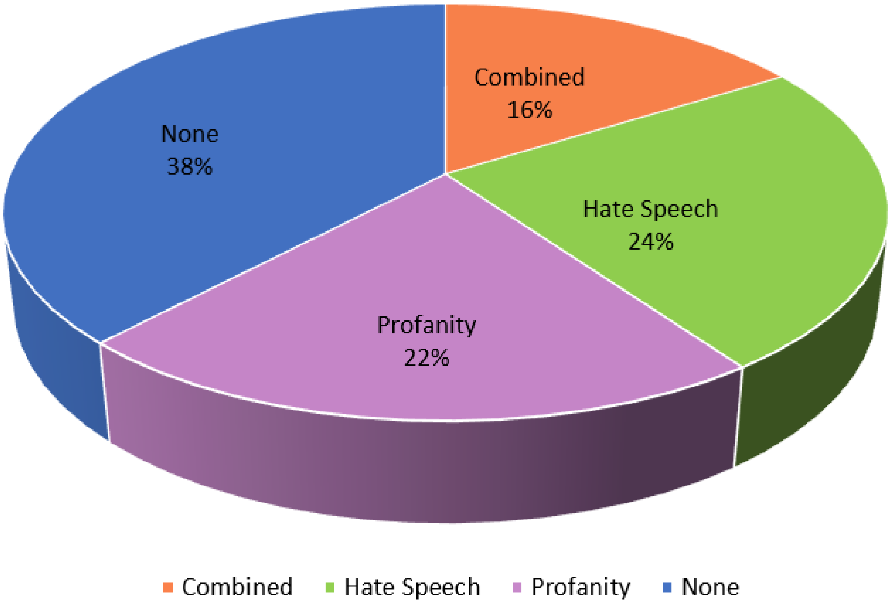


Fig. 4. Total Percentage of Tweets based on the Combination of Hate Speech and Profanity.

In **Fig. 6**, the bar chart shows the proportion of tweets that contain profanity words based on the mark that was manually sorted to decide whether a tweet depicted “self hatred”, “beliefs”, “sexual”, “racist”, “politics”, “bully” or “criticism”. More than half of the total tweets i.e. 57.58% of them portrays criticism towards people. Sexual and bully tweets covered 22.10% with just 1.9% difference between the two. Racist tweets are just

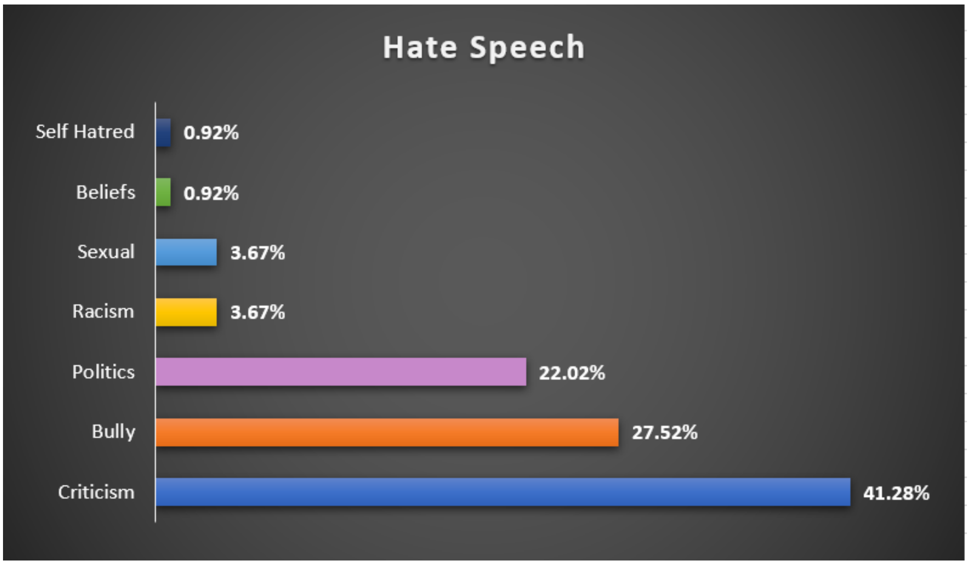


Fig. 5. Total Percentage of Tweets containing Hate Speech based on the Label

1% more than that of self-hatred consuming tweets. And minorities of the tweets attack the beliefs and politics by 2.02% and 1.01% respectively.

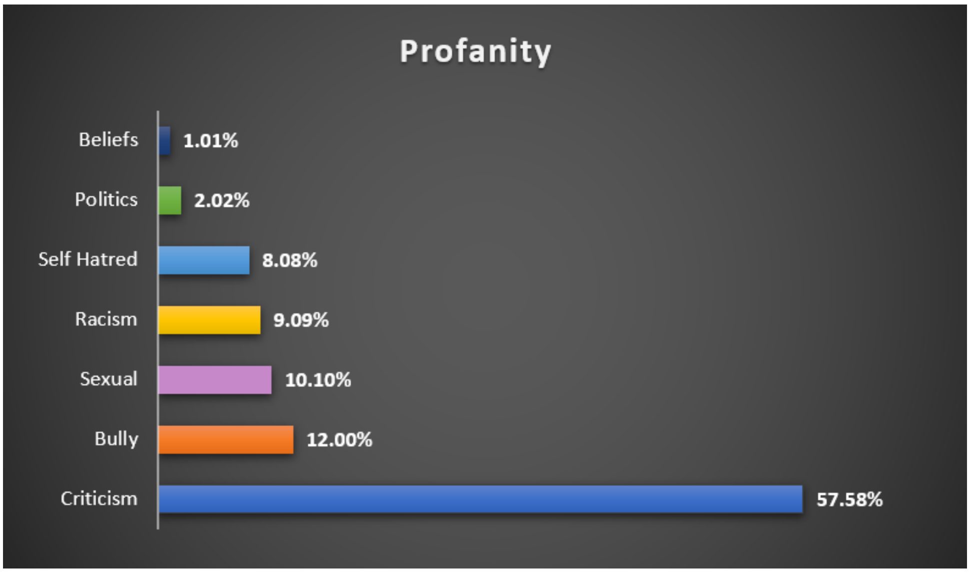


Fig. 6. Total Percentage of Tweets containing Profanity based on the Label

In **Fig. 7**, the bar chart depicts the proportion of tweets that contain both hate speech and profanity based on the label manually assorted to determine if a tweet depicts "self hatred", "beliefs", "sexual", "racist", "politics", "bully" or "criticism". Majority of the tweets containing slang and hatred are bullies i.e. 35.62% of the total. After that, most of the tweets are sexual covering 27.04% of the total tweets. One quarter of the tweets are criticism and attacking towards politics with 12.33% and 13.70% respectively. The percentage of racist tweets comes just after which is 8.22%. A small proportion of the tweets are portraying self-hatred and negativity towards beliefs in 1.37% each.

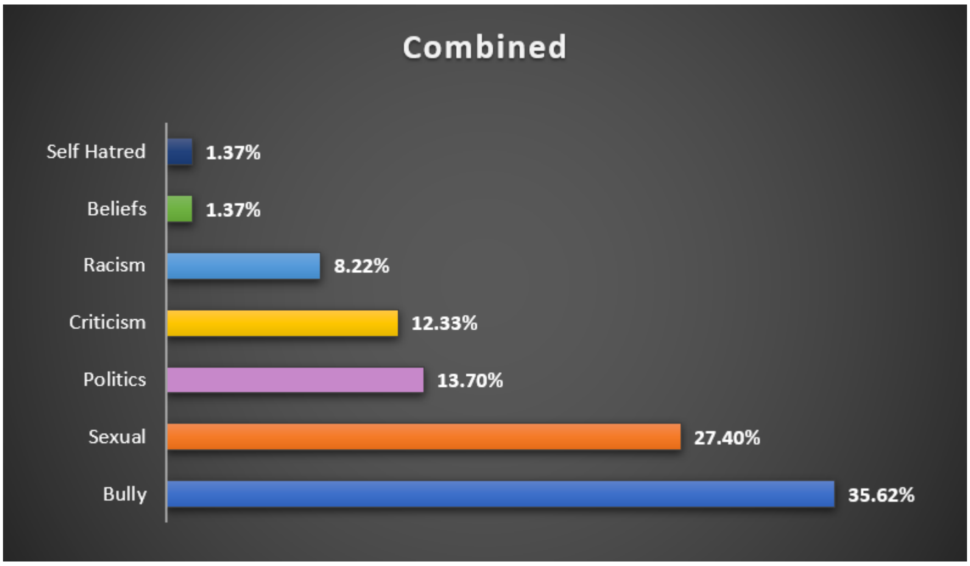


Fig. 7. Total Percentage of Tweets containing both Hate Speech and Profanity based on the Label

In **Fig. 8**, the bar chart illustrates the proportion of friendly tweets which do not contain any negative speeches based on the label manually assorted to determine if a tweet depicts "wishes", "grateful", "inspiring", "praising", and "thoughts". More than 50% of the tweets i.e. 56.17% are the genuine thoughts of people. After that, gradually comes the percentage of praising, wishes, gratefulness and inspiration containing tweets. Almost 30% of the tweets contains praises and wishes and the rest of the 15% tweets consist of inspiration and gratefulness.

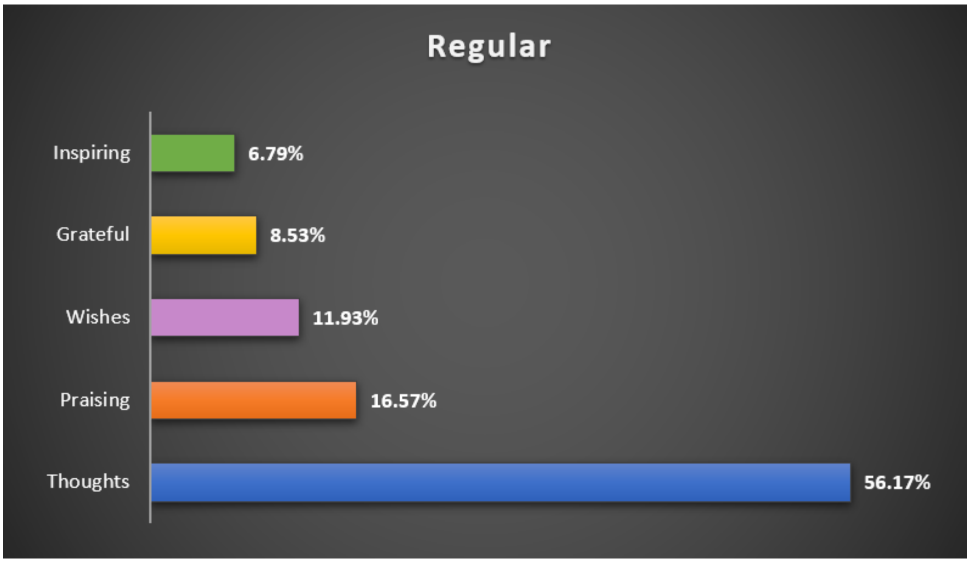


Fig. 8. Total Percentage of Friendly Tweets based on the Label

In **Fig. 9**, the bar chart depicts the proportion of the total tweets depending on the label manually assorted to determine if a tweet depicts "wishes", "grateful", "inspiring", "praising", "thoughts", "self hatred", "beliefs", "sexual", "racist", "politics", "bully" or “criticism”. The green colored bars depicts the friendly tweets and the rest are either hate speech or profane tweets.Almost half of the total tweets i.e. 45.56% of them portrays general thoughts of people. It is clearly illustrated that most of the tweets do not contain hate speech or profanity, a total of 81.15% tweets are friendly and non violating. On the other hand 18.85% tweets are hateful or profane tweets.

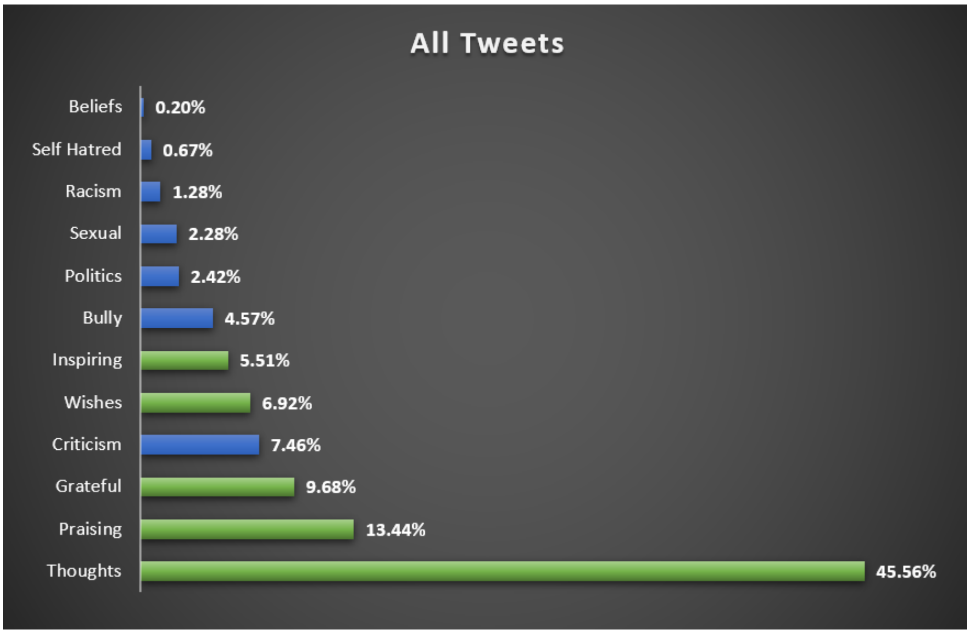


Fig. 9. Total percentage of Tweets based on the Labels

5 Conclusion

Although the dataset is small yet the process of collecting it is a bit different and anonymized. Subcategories of the labeled tweets are helpful for detection, as conflict may arise while differentiating them through machines.