Youtube Comment Analysis for Identifying Cyber Bullying and Hate Comment Likelihood

A study in Sentiment Analysis and Logistic Regression

Shardul Singh
B.Tech. Computer Science Engineering
PES University
Bengaluru, India

Anish Murthy
B.Tech. Computer Science Engineering
PES University
Bengaluru, India

Shreyas V Patil
B.Tech. Computer Science Engineering
PES University
Bengaluru, India

***Abstract*—Presently, most websites use a report system for checking for offensive comments, where users report incidents they feel are offending them and administrators see if they need to intervene. We aim to use comment sentiment analysis to determine which kinds of videos (in terms of tags, genres, likes and dislikes, views) are more likely to contain offensive comments that aim to offend and bully. Several other factors such as polarization of views and types of users will also be taken into consideration to have a more subjective idea of comments being checked.**

Keywords—Sentiment analysis, Logistic regression, Naïve Bayes classifiers

I. Introduction

The modern age of the internet has brought with it's "information revolution" a new way to form communities and make friends. Social media is a revolution in every sense of the word. Being able to not only share information across great distances with millions of people at once, but do so with its current cost, ease and comfort was unimaginable 20 years ago. It is this ease and comfort which played the most important role in bringing social media to its current popularity

The smart phone is probably the next biggest contributor to the importance social media now has in our lives. We are now able to access information to and contact people we want to at any time of the day, without the need of bulky routers and time consuming dial up connections. A small hand held device that every person caries on them all the time, allowing them to capture their moments on film and photo and update statuses at any time allows for a real time sharing of experiences.

The popularity of social media goes beyond it being just "liked". It has become a constant part of our daily lives, to the extent that people feel the need to post their food, daily routine, major life moments, political opinions and even daily experiences on a daily basis. However, with this increase of closeness there has been the emergence of people's "social identities". How we portray ourselves online, the things we comment on and how we behave are an entire dimension to our identities that didn't exist before. These identities range from people just voicing their opinions better than they once did, to slightly uglier forms due to the anonymity they provide.

II. Need to measure Likelihood of Cyber Bullying

The importance of people's online identities in their lives means that they can be hurt online just as well as they can be in real life. This is a fact that people often fail to acknowledge when they interact online, especially when they have anonymous identities. Cyber bullying is a very real and very ignored problem that exists today, and it needs to be tackled as best as it can. People shrug off victims of cyber bullying by simply stating that it's not real and that they should just get over it. However, since people often pour their hearts into their social identities, we believe cyber bullying has a huge psychological impact on its victims.

Several studies have shown direct links between cyber bullying and depression, especially amongst teens and other people who spend a lot of time online. The main problem is that the internet and social media are a very huge part of our lives and both exist ubiquitously and have deep impact on our lives. All of this given, we feel that they have very little policing and very few laws simply because people haven't realized that their online interactions need to follow norms and need to be monitored just like real life interactions should be. The degree of monitoring we are referring to is just enough so that these websites can become a safe enough environment for one to at least expect a discrimination and racism free environment, one where they don't need to live under a burden of fat shaming or name calling, and one where they can safely have positive interactions with other humans without fear of psychological trauma, even in the slightest sense.

On platforms like YouTube, which allows artistes and social butterflies alike to share videos of their daily lives, the mask of anonymity often leads to a lot of bullying in the form of discouraging comments, hate speech, racism etc. Since a lot of people, under the veil of anonymity, don't correctly identify bullying as bullying, we feel it is essential to not rely entirely on the "report system" to check offensive content and instead build our own model to determine content that is more likely to garner hate speech and bullying. After identifying problem areas better, we feel it'll be easier for administration to monitor comment sections on those specific problem areas better.

III. Sentiment Analysis: An Overview

Sentiment analysis or opinion mining is a discipline of natural language processing that works to determine the emotional tone behind a series of words. It's generally used in blogs, comments and threads to determine the attitudes, opinions and emotions expressed within an online mention. It's a powerful tool in the present day in age, where most data is unstructured and available in varied forms, generally with bad grammar, sarcasm etc. The biggest issue with sentiment analysis and text processing, thus generally is in the initial setup of the model. Since data is so unstructured, it takes a long time to set up each manual parameter and rule in order to eventually accurately judge the sentiment behind a statement. Even so, it can often not detect incorrectly written words and sarcastic comments. In order to minimize some of these shortcomings, analyses are done in several layers:

A. Dictionaries

This is the first and most basic filter applied to words. Put simply, words are analysed independently of each other after removing common language filler words (eg but, and, a). It requires a preprocessing of words in order to normalize them first (eg nice, niice and niiiiiice all need to be normalized to the word nice). The words are then matched against a pre defined dictionary of words which have been categorized according to their sentiment (eg curse words are read as having a mean sentiment, words like "good" are seen as positive comments). Thus, comments are rated simply on the occurrence of these pre defined lists of words. For our purposes, several such dictionaries are available freely online in sources listed below. Such sources have been consulted to build a large dictionary.

B. Contextual analysis

This is a much broader and vaguer level of analysis, and includes testing which words come with other words, and other things like how polarized is the current videos (likes to dislikes ratio), what tags are associated with the video etc. All such contexts are much tougher to establish and require a larger amount of setup.

IV. Previous work

Following is a list of previous work done on similar themes. Each paper has a short description.

A. Properties,Prediction,andPrevalenceofUsefulUser-Generated Comments for Descriptive Annotation of Social Media Objects, By : ElahehMomeni, Claire Cardie and Myle Ott

1) Summary:

Two levels of text sentiment analysis were done to determine the usefulness of comments, classified into Text based Linguistic features (TL)(sentence length, readability, number of punctuation marks etc) and Semantic and Topical Features (ST) (polarization, tags, context etc). After splitting the data into testing and training datasets, two experimental classifier models were set up namely a logistic regression model and a naïve bayes classifier model. Then, they experimented with several groups/ subgroups such as genres, polarization levels, popularity and videos tags to find out how the classifier model varies in different subgroups and created niche models for each.

2) Results

Several influencors were found for the useful. For example topics related to users have lesser useful comments than those related to events. Relative number of useful comments also seemed to decrease the more polarized the video was.

B. Mean Birds: Detecting Aggression and Bullying on Twitter: By Despoina Chatzakou, Nicolas Kourtellis ,Jeremy Blackburn Emiliano De Cristofaro , Gianluca Stringhini, Athena Vakali .

1) Summary

After an initial cleaning of text and removal of filler English words, the data is categorized into bins using sentiment analysis techniques. This twitter comment data was then judged on several metrics: number of hash tags used, number of times the user tweets etc as well as the popularity of a user, the amount of replies they get etc. Three kinds of classifiers were set up, using tree based and ensemble classifiers followed by 10 fold cross validation.

2) Results

Aggressive users show similar behavior to spammers in terms of followers, tweet frequency etc. Text based features were found to have little impact on the aggression level of a user, and network based parameters such as tweet frequency were much more significant. They achieved an accuracy of 91%

V. Problem Statement

We are attempting to improve upon previous work down through a rigorous analysis of sentiments with several tags and parameters associated with a YouTube video with what we believe to be the right metrics. Since account/ user banning is difficult in the anonymity that exists and fake accounts just pop up every day, we aim to curb down on hate comments themselves. Also, we hope to create a cyber bullying index or rating in the end, a numerical value which determines the risk level of hate comments on a given video. Such an index can make administration and moderations on these websites much simpler since high risk videos, based on metrics, can be monitored with more care. We believe the creation of this index is entirely unique.

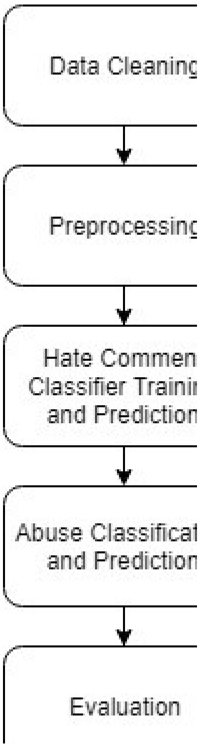
In attempting to obtain this index, we are using the dataset, Trending YouTube Video Statistics and Comments dataset by Mitchell J, which lists the data from trending videos in the US and UK collected over the period of one month. This data includes the video title, channel title, category, tags, views, likes, dislikes, date and data on the most relevant comments for each video. The dataset contains all fields necessary for the purpose of comment and trend analysis our project requires and hence we have chosen to use it in our project. We are using sentiment analysis on the comments to obtain the ratio of hate comments per category of videos as well as ratio of the likes to hate comments for each video.

We are also using a dataset of common abuse words in order to obtain the amount of abuse each video obtains. We then use this data to obtain information on the amount of abuse each genre receives.

Thirdly, we use a dataset called the Hate Speech Identification by crowdflower which lists twitter tweets manually classified into the categories, hate speech or not hate speech, along with more data on the confidence of this classification, the presence of abuse and the tweet text. We have used this dataset in order to train the classifier for detecting hate speech which is carried over into the YouTube dataset.

On attempting this project, we find that there is a potent lack of relevant datasets for hate speech detection with respect to YouTube, as well as datasets of sufficient size; hence we have used a twitter hate dataset for comment classification. We have also found that the datasets tend to have a large number of special characters, which makes cleaning the dataset difficult. The comments in the youtube data also tends to have many spelling errors or slangs, which make the classification difficult.

VI. PROCEDURE



A. *Cleaning the data:*

The data we are using has many special characters which led to errors in importing the data; this is cleaned by checking the video id's for valid values, which we use to filter the data.

We then tokenize the dataset and remove stopwords in order to obtain the usable form of the comment texts. This is achieved using the quantida package in R.

B. *Pre-processing:*

In order to facilitate further processing, we have created several extra parameters for our YouTube data. We have added the slang frequency, abuse ratio, polarity and abuse level as separate columns in the data frame.

The slang frequency is calculated by summing up the number of abusive terms in each comment for each video. The abuse ratio is found by calculating the ratio of abuse to the number of comments in each video. The polarity is the average of the ratio of dislikes to likes for each comment in the video. The abuse level stores the categorization of the magnitude of abusive comments for each video.

We have also attempted to classify the comments into 'hate comment' and 'not hate comment' using a separate classifier run on the twitter hate speech dataset, which we have added as a separate column.

1) *Slang Frequency:*

We have used the abuse dataset in order to classify words in each comment of the dataset as 'abusive' or not. This data is used to find the total slang frequency of each video and hence the slang frequency of the categories.

2) *Abuse Ratio:*

We use the previously obtained Slang Frequency data to obtain a ratio of the amount of abuse received per video to the number of comments a video receives in total.

3) *Polarity:*

We calculate the ratio of dislikes to the total number of likes and dislikes each comment has received. We then use the average of this values for each video in order to obtain the average polarity a video receives. We use the same to obtain the same for each category.

4) *Hate speech categorisation:*

We have used a Naïve Bayes Text Classifier to obtain the classification for each comment. The training of this classifier was carried out on the Twitter Hate speech dataset, which resulted in a classifier with an accuracy of 84% on that dataset. The classifier is then applied to each comment of the YouTube dataset. This data is used to find the Hate Speech ratio for each video and hence each category.

5) *Creating levels for hate speech and abuse:*

We have created 4 levels or bins of the independent, based on the four quartiles of distribution. Therefore an abuse level of 1 signifies the least level of likelihood of abuses and a level of 4 signifies the highest likelihood. Similarly, for hate speech 1 signifies least likelihood of hate speech and 4 signifies the most.

C. Prediction:

We have used Logistic Regression, Naive Bayes classification, decision trees and SVM to attempt to predict the likelihood of a new video being uploaded to YouTube getting a) hate speech and b) abusive comments. We therefore have built for classification models each for the prediction of two separate independent variables.

1) Paritioning the data:

To obtain a usable training and testing set, we have partitioned the data in a ratio of 70:30 for train and test data. We have then created separate dataframes for each containing only the relevant data we shall be using

2) Building the model:

Quite simply, 8 models were built (4 models for 2 independent variables each) in order to categorize data into 4 bins ranging from 1 to 4. The model was trained on the earlier partitioned training dataset for each of the 4 models.

D. Evaluation Criteria:

We have used 2 basic criteria for evaluating the success of each model.

1) Accuracy:

We tested our predicted levels vs the actual levels and attempted to see how our model fared against real data. Since likelihood of error is much higher on a model that attempts a 4 way classification, the original accuracy obtained was extremely low. Hence a modification was made to the accuracy criteria, and a classification error of 1 was allowed. Therefore, if the actual likely abuse level of a video was 2, a prediction of 1,2 or 3 was considered valid. This accuracy criteria works well since it still accurately predicts our highest abuse/hate level i.e. 4 accurately, while cutting down on false negatives by incorporating predictions of level 3. So false positives may increase, but have a much lesser consequence than any increase in false negatives.

2) Higher level accuracy or sensitivity:

This is a metric especially designed by us to be able to analyze the sensitivity of predicting the higher levels correctly. It looks at the probability of classifying a level 4 threat as either a level 3 or a level 4 threat correctly. This metric will henceforth be referred to as H_Accuracy. This is important, perhaps more so than the accuracy since our filter's main aim is to identify higher bullying threats accurately, and classification of lower threats doesn't really matter much.

VII. Experimental results

A. Visualization inferences

We first did an analysis of the data to find any interesting insights before building any models. We wished to find out the relationship between likes and genres, as well as dislikes and genres.

An analysis of likes vs genres yielded expected results, with music getting the most number of likes (possibly because of getting the most number of views as well) and Howto and Style getting the least number of likes. When attempting to find out which genre gets the most dislikes, we assumed it would be politics or perhaps entertainment. We were surprised to find comedy having by far the most number of dislikes, possibly indicating people's strong sense of opinion when it comes to entertainment, or perhaps a strong sentiment against abusive and racist comedy, as is the trend on television and the internet these days. Extending this idea, we wished to see the degree of polarization between different genres. We expected to find politics as the forerunner, due to the inherent polarized nature of politics. We we re again pleasantly surprised to find Howto and style overtaking politics as the most polarized genre. This probably indicates people search for problem solutions on youtube all the time, and like/dislike them quite often based on how helpful it was, without considering that it just wasn't useful to their specific case.

Next, we wished to determine whether there was a correlation between genre and abuseratios. There seemed to be a correlation, just as would be intuitively evident, with news and politics getting the most abuse and education getting the least. Other academic fields like science and technology very similarly had very low abuse ratios indicating that bullies are probably fuelled by aggressively pushing their opinions of arts and politics on people, rather than just a random urge to bully. This is supported by the fact that genres such as comedy, entartainment and news and politics get far more abuse proportionately.

Next, we wanted to compare our two evaluation criteria I.e. abuse ratio as well as hate comment classification to see how close their results are to each other.

We note that a plot of hate comment ratio vs genre gives a much similar result as that of abuse vs genre, with certain changes. While Howto &Style shows the most hate comments, education is now the second least threatened genre overtaken by nonprofits and activism. We clearly see that this criteria shows certain deviations from the first one. While we don't have a solid idea of which criteria is likely the best one to judge our results on, we believe it's the right choice to move forward with both, since both have different merits as will be discussed shortly.

B. Model building and analysis

TABLE I. PREDICTION OF ABUSE LEVELS

	Accuracy	H_Accuracy
Logistic Regression	0.6941432	0.4490239
Naive Bayes Classification	0.5986985	0.132321
Decision Tree	0.7483731	0.4229935
Support Vector Machine	0.7114967	0.4880694

TABLE II. PREDICTION OF HATE SPEECH

	Accuracy	H_Accuracy
Logistic Regression	0.9587852	0.537961
Naive Bayes Classification	0.6225597	0.06724512
Decision Tree	0.9045553	0.4295011
Support Vector Machine	0.9305857	0.6138829

Based on abuse levels, the best accuracy is obtained by using decision trees. The problem with this, however, is that accuracy is our main criterion here. H_Accuracy has much more significance in this case since our main aim from the very start has been to be able to classify high level threats of bullying properly. Thus, an SVM is best suited due to a good balance of accuracy and the best H_Accuracy possible. However, our results for this independent variables are abysmal at best, a 48% sensitivity isn't very good. This leads us to believe that either a) it is not possible to use other predictors to estimate the amount of bullying on a data or b) Our approach to bullying ie the number abuse words is wrong. Since abuse has become a common part of today's youth, this may not be the ebst indicator of bullying in the first place. Simply having more abuses in the comments could be an indicator of a younger viewership rather than an indicator of bullying in general.

Our second independent variable both gives more promising results, and seems like a better embodiment of cyber bullying. By training on a large, curated dataset of hate comments from twitter, we tested our data and classified it as hate speech or not hate speech. We then built our models to check whether non-textual parameters can be used to mirror this output. Our results were a resounding success, with SVM giving the overall best results. While Logistic regression certainly gave a better accuracy than SVMs, SVM has by far the best H_Accuracy. As discussed earlier, this will be our primary criterion for evaluating a model.

VIII. CONCLUSION

Having trained on a large dataset of twitter hate comments, we believe our second analysis model is much more accurate. With that in mind, we believe the SVM created is far superior to all the other methods of prediction and yields decent enough accuracy for us to be able to do the task we had set out to: to reduce the text processing load on youtube bullying filters by doing an initial check on easier to process parameters such as likes/ dislikes and genres.

As an improvement to this project, we would also like to work on tag analysis in the future to be able to further classify genres into sub genres. We believe the viewership of a video defines the kind of comments it'll get, and tag analysis is crucial for this. However, a much larger dataset needs to be analysed for that, for which we currently lack the processing power.

IX. Acknowledgment

This project has been undertaken under Dr. Gowri Srinivasan for partial fulfillment of the requirements of the course “Data Analytics”, course code UE15CS322 and has been made possible due to the support of the PES University faculty.

References

[1] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, GianlucaStringhini, Athena Vakali, “Mean Birds: Detecting Aggression and Bullying on Twitter” arXiv:1702.06877v3 [cs.CY] 12 May 2017

[2] Elaheh Momeni, Claire Cardie and Myle Ott. “Properties, Prediction, and Prevalence of Useful User-Generated Comments for Descriptive Annotation of Social Media Objects”:

[3] Hsu, C.-F.; Khabiri, E.; and Caverlee, J. 2009. Ranking comments on the social web. In Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04, CSE ’09, 90–97. Washington, DC, USA: IEEE Computer Society.

[4] Sigurbj¨ornsson, B., and van Zwol, R. 2008. Flickr tag recommendation based on collective knowledge. In Proceedings of the 17th international conference on World Wide Web, WWW ’08. ACM.

[5] https://blog.medallia.com/featured/text-analytics-works/

X. Role of Each Member

Shreyas V Patil: Preprocessing and feature engineering, visualization

Shardul Singh: Model building and model analysis

Anish V Murthy: Preprocessing, Text classifier for prediction(Hate speech classifier) and video editing