Sentiment Analysis on Twitter and YouTube

Hemanth Reddy Karri
hkarri1@binghamton.edu
Binghamton University
Binghamton, New York, USA

Punit Paresh Jagani
pjagani1@binghamton.edu
Binghamton University
Binghamton, New York, USA

Akhil Parimi
aparimi1@binghamton.edu
Binghamton University
Binghamton, New York, USA

Vijay Kumar Kadamanchi
vkadama1@binghamton.edu
Binghamton University
Binghamton, New York, USA

ABSTRACT

Toxic comments and personal attacks have become increasingly common on social media platforms, online news commenting areas, and many other public venues on the Internet. However, deciding whether or not to "flag" a comment or post is complex and time-consuming. Not only would automating the process of detecting abuse in comments save website moderators time, but it would also promote user safety and improve online discussions. We're particularly interested in tweets and YouTube comments that contain any harsh or toxic language. We want to identify tweets and YouTube comments that contain toxic phrases and label them as negative using the data we've acquired in a certain time frame.

KEYWORDS

Machine learning models for classification, text mining, text analysis, data analysis, data visualization, MongoDB, Twitter API, YouTube API

ACM Reference Format:
Hemanth Reddy Karri, Punit Paresh Jagani, Akhil Parimi, and Vijay Kumar Kadamanchi. 2018. Sentiment Analysis on Twitter and YouTube. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The issue of trolls and spammers is getting more prominent as discussions move more and more to internet forums. Manually moderating comments and discussion forums is time-consuming, and organizations are compelled to use contractual or outside moderators to deal with the high volume of comments. We're looking for hate speech on Twitter and YouTube. We consider a tweet or a comment to be hate speech if it incorporates racist or sexist comments. The main objective is to classify negative tweets and comments from other tweets and comments.We'll be collecting real world data from popular media channels like Twitter and YouTube for this project. We'll use the data from these two sources with a defined time frame to discover and highlight any negative sensing content. We also aim to visualize the classified data of Twitter and YouTube.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

2 RESEARCH QUESTIONS

1) How much of the data is positive?

2) How much of it is negative or offensive?

3) How much of the data is neutral?

We categorized all of the data into three categories: positive, neutral, and negative sentiments, and represented them in bar graphs. We can clearly see and comprehend the proportion of negative comments to the overall number of comments or tweets.

3 DATASET

We have collected our data from Twitter and YouTube. We did not employ any filters and collected all of the data without any sort of pre-screening.

3.1 Data Sources

Below are the sources:

DataSource 1: Twitter- We use TweetStream python module that can be used to get tweets from Twitter's streaming API.

DataSource 2: YouTube- The Google API Client Library for Python is designed for Python client-applications. Using this library, we can get all of the comments posted of random YouTube videos.

4 DASHBOARD ANALYSIS

Dashboard will be built using HTML and Bootsrap 5 and it will be hosted as a web application.

We will implement two date pickers, which will enable us to set a start and end date. We can then show the visualized bar graphs for those date ranges. This analysis will display the distribution of various sentiments, such as:
(1) Negative Sentiment
(2) Positive Sentiment
(3) Neutral Sentiment

5 CONCLUSION

This dashboard will be shown as a website. When we examine the created graphs, we will find answers to all of our research questions. Two graphs would be created. One for Twitter and one for YouTube. We may choose a date range for which the analysis will be displayed, and the analysis will only be displayed for those dates. It would be quite simple to comprehend, and we would be able to quickly determine how many tweets are negative, positive, or neutral in nature.