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, 1 page. 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.

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?

3 METHODOLOGY

We utilized the TweetStream python package to gather random tweets from Twitter and GoogleAPIClient to extract comments from YouTube, as indicated in the Data Acquisition section. We saved all of the collected data in MongoDB, from which we will get the data for analysis.

We received an average of 7 million tweets and over 1.8 million YouTube comments over the course of seven days. We have the ability to use every single comment and tweet without screening them since we are not aiming to confine our data to a certain topic.

The data is then extracted from MongoDB and cleaned using a regex function. This is how we remove all of the noise from the data. This function, for example, can be used to eliminate tweet content written in a language other than English.

We then use the clean data to perform sentiment analysis to find out if the particular tweet/comment is Positiive, Negative or Neutral. Once the analysis of the all the data is complete, we can visualize the data in the form of bar graphs.

4 CONCLUSION

Finally, our goal is to do data analysis on the data acquired from Twitter and YouTube following the data collection process. We'll use descriptive and qualitative analysis to answer the research questions. Using data visualization tools, we will also graphically depict the outcomes and trends.

Conference'17, July 2017, Washington, DC, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456