Twitter Sentiment Analysis to improve Peace and Justice

Amogh Avadhani Gaurav Gopalkrishna Qianyu Zhao

***Abstract*— Effective inspection of data from social media can help in analyzing trends in the communities, including communal sentiment. Exploiting different aspects of social media trends can reveal public opinion which can be utilized by governments and law makers for efficacious administration.**

I. INTRODUCTION

Social media can be a powerful tool to be utilized in analyzing user sentiment. Users express their concerns and opinions on social media which can be studied to analyze trends in the community. Twitter provides a platform where people express their views regarding concerned matters easily, the data from which is suitable and can be leveraged for analyzing public sentiment.

A. Sustainable Development Goal 16: Promote just, peaceful and inclusive societies

Goal 16 of the Sustainable Development Goals is dedicated to the promotion of peaceful and inclusive societies for sustainable development, the provision of access to justice for all, and building effective, accountable institutions at all levels.

The targets of SDG 16 are:

- Significantly reduce all forms of violence and related death rates everywhere.
- End abuse, exploitation, trafficking and all forms of violence against and torture of children.
- Promote the rule of law at the national and international levels and ensure equal access to justice for all.
- By 2030, significantly reduce illicit financial and arms flows, strengthen the recovery and return of stolen assets and combat all forms of organized crime.
- Substantially reduce corruption and bribery in all their forms.
- Develop effective, accountable and transparent institutions at all levels.
- Ensure responsive, inclusive, participatory and representative decision-making at all levels.
- Broaden and strengthen the participation of developing countries in the institutions of global governance.
- By 2030, provide legal identity for all, including birth registration.
- Ensure public access to information and protect fundamental freedoms, in accordance with national legislation and international agreement.
- Strengthen relevant national institutions, including through international cooperation, for building capacity at all levels, particularly in developing countries, to prevent violence and combat terrorism and crime.
- Promote and enforce non-discriminatory laws and policies for sustainable development.

In analyzing tweet impressions, a correlation could be found between civilian sentiment and the effective peace and justice in the region. The sentiment analysis tool can then be employed by law enforcers around the world to analyze peace and public safety, enabling them to undertake judicious decisions.

The entire pipeline uses Apache Spark with MLib machine learning library.

II. DATA COLLECTION

A multi-faceted approach was employed for data collection. Collection of data had many unforeseen challenges. The challenges arose chiefly due to Twitter's usage and anti-crawler policy and untrustworthy data.

- **Streaming APIs:** Twitters streaming APIs was initially used to collect data. Twitter streaming was made easier by using Tweepy and Twython python library Twitters usage policy includes rate restrictions, and also the data streamed is only recent data. The analysis in the project demanded historical data over a large period of time in order to discover public sentiment. Therefore, an approach which could circumvent these policies had to be put in place.
- **Crwaler:** Scrapy framework was used to crawl twitter web pages via twitter search. User action was simulated by keyword search on twitter.com/search. XPath was used to decode the html text in the web pages.

Two separate sets of data was collected, a training set on which the Naive Bayes classifier was trained to create a model, and a testing set for prediction.

- **Training set:** The training set consisted of pseudo labelled data. This was achieved by collecting tweets data related to certain hashtags which are related to safety, peace and security. A point to note here is that our model will need instances of data which have both positive and negative tweets data related to safety, peace and security. Therefore, the hashtags on which the data is based has negative connotations, positive connotations and neutral connotations with regards to safety, peace and security.
  Positive hashtags: #faith, #humanity, #onelove, #firefighter, #veteran, #heroes, #marine, #vet, #thank, #grateful, #nypd, #lapd
  Negative hashtags: #injustice, #corruption, #discrimination, #racism, #war, #abuse, #NotOneMore, #guncontrol, #gunsense, #MeToo, #gunviolence,

#secondamendment
Neutral hashtags: #sports, #NFL, #Got, #datascience, #hollywood, #blackfriday

- **Testing set:** The testing set consists of data from different sources. A certain amount of tweets were collected using the streaming APIs. During this process, lot of metadata related to the tweets like username was also collected. These usernames were then used as keywords in the search while crawling. Another way usernames were collected is via tracking the followers of many popular twitter followers.

A. Summary of data used

Training data: approx. 1GB of twitter data
Test data: approx. 7GB of twitter data

III. Preprocessor and tokenizer

Apache spark is used as a data pipeline to hold dataframes loaded from the text files in which data is collected. The collected data is preprocessed to remove unwanted features.

- Filtering english words
  Twitter is a global platform supporting many languages across the world, which means that we end up with tweets with different languages. Also, users can be multilingual and write tweets with mixed languages. Since we are interested in only the US related analyses, it is necessary to filter out non English words from the data. The collected tweets consisted up to 70% of English data.
- Nltk tokenizer
  Nltk tokenizer was used to tokenize the cleaned data. The stop words such as 'is', 'an', 'the', which do not have an effect on the overall tone of the tweet was removed using nltk.
- Removing non important data from tweets
  Strings such as URLs, punctuations, numbers, @mentions etc are not relevant in our studies and were removed.
- Lemmatisation
  Inflected forms of words are grouped together to a single item. For example: good, gud, goooood etc. are all mapped to a single token good.
- Duplicates Any duplicate data in the training set which can adversely impact the models is removed.

After all the above preprocessing, we end up with roughly 70% data as testing data and 30% of the data as training data.

IV. Classifier and Sentiment Analyzer

A. Pipeline

The entire pipeline uses Apache Spark with MLib machine learning library. MLib fits into spark's API and supports many of Python's libraries. MLib is easy to plug into Hadoopś workflow and is optimized to provide better performance gains.

B. TF-IDF

The measure to evaluate the importance of words in the document was calculated using TF-IDF. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

1) TF: Term Frequency measures how frequently a term occurs in a document. The term frequency is often divided by the the total number of terms in the document.

2) IDF: Inverse Document Frequency measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones. IDF is calculated by the log of the value attained by estimating the total number of documents divided by the number of documents containing that particular term.

C. Naive Bayes classifier

Considering the fact that Naive Bayes performs well with discrete classification, we decided to go forward with Naive Bayes.
The Naive Bayesian classifier is based on Bayes theorem with independence assumptions between predictors.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

D. Sentiment analysis

The Naive Bayes classifier is trained on the pseudo labelled data (on hashtags) and a model is generated. The Naive Bayes model is used to classify the tweets in the test data. Python's TextBlob library is used for analyzing tweet sentiment.

E. Two fold cross validation

To estimate how accurately the Naive Bayes predictive model will perform, 2-fold cross validation was carried out which resulted in an accuracy score of 0.9122 for the model.

V. Results

After running the classifier on the test data, we can infer a good idea about the public sentiment in a region about topics related to peace and safety. We can correlate this to say how safe people feel in their respective city, county or state.

References

[1] https://scrapy.org/
[2] https://spark.apache.org/docs/latest/ml-guide.html
[3] http://textblob.readthedocs.io/en/dev/
[4] https://www.programmableweb.com/api/twitter
[5] http://www.tweepy.org/
[6] http://www.nltk.org/

Fig. 1. Pipeline