Analysis of Sentiments in Twitter

Ziyang Yang
University of Colorado Boulder
ziya6127@colorado.edu

Nicholas Montoya
University of Colorado Boulder
nimo0295@colorado.edu

Zihan Zhou
University of Colorado Boulder
zizh1583@colorado.edu

1 PROBLEM STATEMENT AND MOTIVATION

Nowadays, social media have tons of users, and millions of people share opinions on different aspects of life every day. It's interesting to figure out how people express their sentiments through different words, emojis, etc. This paper and research serves to provide a better understanding about human sentiment on social media, especially in Twitter. Also, we want to find the emotions behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention. Our project will serve as a study of sentiment analysis based on Twitter data. The applications of sentiment analysis are broad and powerful. For example, the shifts in sentiment on social media have been shown to correlate with shifts in the stock market. In this project, the research will draw motivation from people to understand and analyze the people's opinions more efficiently.

2 DATASET SUMMARY

We are going to use several datasets in this project:

- **Sentiment140 dataset**

It contains 1.6 million tweets extracted using the twitter API. These tweets are also pre-tagged with sentiment. Also, the insight that can be gained from millions of Tweets will overshadow the concerns about reliability of a single Tweet. https://www.kaggle.com/kazanova/sentiment140/data

- **USA: Geolocated Twitter Dataset**

Contains 200,000 tweets in total from April 2016. This set comes with multiple files that also contain information the popularity of tweets based on retweets, favorites, and a combination of both. There are also location details however, we will not be using that information.

3 PREVIOUS WORK

There has been a lot of work done in both sentiment analyses along with predictive language models. Regarding sentiment analysis, one of the things we needed to look at was tokenization. This is especially important for special characters, like emojis, that are relevant to the text but not easily classified. We used the article found at the following link: [1]
http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2857/3251?height%3D90%%26iframe%3Dtrue%26width%3D90%

However, they also used a 5-point system to score sentiment. We will be using a simpler 3-point system. They also tracked hashtags meticulously. This will not be used in our project. The main thing taken from this source is the use of tokenizing characters.

Like mentioned before, there is an article which talks about the sentiment analysis and opinion mining. From the article "Twitter as a Corpus for Sentiment Analysis and Opinion Mining"[2], the author shows how to automatically collect a corpus

[1] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," ICWSM, vol. 11, pp. 538-541, 201

[2] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS, Bˆatiment 508, F-91405 Orsay Cedex, France.

for sentiment analysis and opinion mining purposes. Also, they perform linguistic analysis of the collected corpus and explain discovered phenomena. This can lead us to better understand the sentiment analysis and help us to achieve our goals more efficiently.

There is an article which talks about the challenge and application of the sentiment analysis and opinion mining. In the "Application of Sentiment Analysis"[3], the author quoted Pang-Lee's idea of application of different category such as application to certain Review-Related websites and certain business in order to track customers' tendency. Besides the application, he also talks about the challenge which remind us potential challenge and problems. For instance, sentiment analysis approaches aim to extract positive and negative sentiment bearing words from a text and classify the text as positive, negative or else objective if it cannot find any sentiment bearing words. Also, there is sometime the application cannot understand, like "I forget to turn in my homework. Nice!" Since it has the word 'Nice' in positive, the application will assume it is a positive post, but it is not. We need to find a way to figure out this problem and this article can help us a lot.

4 PROPOSED WORK

The data set has to be cleaned and preprocessed at the beginning. In order to get a clean, useful data sets, removal of unrelated data and meaningless message is required. To accomplish this, there are several methods listed below:

- Most of the tweets contain mentioning somebody else (Ex. @somebody). However, these are irrelevant information and have to be removed in order to get a meaningful data set.

- Some of tweets have one or several links to other website. Although these links may have some relationship with tweet itself, most of them are irrelevant and should be removed.

- Characters such as emojis need to be converted into NLTK characters and interpreted

- The attribute of query should be removed since all of the objects is no query and query is not relevant with the research.

After cleaning the data, we will perform sentiment analysis on the tweets. This will involve multiple steps:

- Separating data by day
  - Separating further by time of day
- Tokenize words in the set being analyzed
- Score each tweet
  - This is where our sentiment analysis algorithm will work
- Compute and store aggregate score

After we have analyzed the full dataset and have scored the sentiment of each tweet, we will use try to interpret the sentiment as it changes throughout the day and over time. Furthermore, we will then train a pre-existing Recurrent Neural Network (https://github.com/minimaxir/textgenrnn) on our dataset and generate tweets.

5 TOOLS

We will use variety of tools to help us work on the analysis, the programming tools to be used are listed below.

- Python: A high level programming language with variety useful libraries and API.
- Pandas: A python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. This library can help the other libraries operate data sets easily.

[3] Subhabrata Mukherjee, Dr. Pushpak Bhattacharyya, "Sentiment Analysis – A Literature Survey," June 29, 2012. Indian Institute of Technology, Bombay.

- Numpy: A python package used for numerical computation on datasets and other mathematical problems. Numpy is a helpful tool for computing large data sets and it is easy to access.
- Scipy: A python-based library for scientific computing. Scipy can help for numerical integration and optimization.
- Tweepy: Tweepy is a python-based easy-to-use Twitter API. Since the analysis is based on twitter, Tweepy is an essential tool for collecting certain account's twitter and their replies.
- Matplotlib: Matplotlib is a python library which produces publication quality figures from variety environments and data sets. Matplotlib is an essential tool for plotting and visualization of certain data sets.
- NLTK (Natural Language ToolKit): This is a library that provides methods for interacting with human language. Instead of fully writing custom language processing methods, we will use NLTK to easily form trees from the words in tweets

In addition to these essential libraries and programming language, there are also some tools to help us collecting datasets from other source and tools for version control.

- Github: Github is a web-based hosting service for version control using git.
- NY times API: NY times API allows us to access New York Times data for analysis. (possibly)

6 MILESTONES

The main milestones are listed below:

1. Clean data from the Geolocation based tweets for compatibility by March 20rd
2. Perform sentiment analysis on cleaned (geolocated) tweets by March 26th
3. Combine the datasets to produce a singular, larger set that has been cleaned by April 2rd
4. Write code to get tweets by day by April 9th
5. Separate tweets into smaller time slots within a singular day by April 13th
6. Perform analysis by April 19th
   a. Examine sentiment change
   b. Examine frequency of words within similar-sentiment tweets
7. Create model to output predictive tweets based on the results of analysis by April 23th
8. Test accuracy and fine tune all analysis by April 25th
9. Presentation ready by April 27th

7 PEER REVIEW FEEDBACK

During the Peer Review Session, there are several constructively ideas could apply to the research. For instance, proposed work is required even the data sets is download from certain data sets website and has been processed before. Ignorance of preprocessing data sets will cause several serious mistakes and these mistakes may not be found at the end. Preprocessing and cleaning data is required for this project. Besides the preprocessing data, collaborate or communicate with groups that have similar topic is also a good advice. Groups with similar topic may face similar problems and communication between these groups can help solving most of the problems they met.