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. We are going to test the use time of positive and negative words in the twitter posts to explore the relationship between the word and true emotion.

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** [1]

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.

- **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.

Twitter Sentiment Analysis Words Collection

Contains positive and negative opinion words (sentiment words). Also it has many misspelled words in the list. However, these misspelled words appear frequently in social media content. We are going to use these words to analysis and screen the datasets which after combination.

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

1 http://crowdsourcing-class.org/assignments/downloads/pak-paroubek.pdf

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

2Alexander 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.

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 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.
http://crowdsourcing-class.org/assignments/downloads/pak-paroubek.pdf

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.
https://arxiv.org/pdf/1304.4520.pdf

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 tweets have neutral opinions, which is not very useful for our research and analysis. Therefore, we will drop the useless data to keep the accuracy, and at the same time make the complexity become lower.

- 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.

- The data should be screened, since there are lots of different topics in the dataset. We need to screen several datasets and

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

make it possible to combine all of them in the future analysis.

- There are some interrupting contains in the dataset. For example, people like using irony to express their dissatisfactions about things happened. However, it will be really hard to figure out which one is irony, and which one is not. We are going to create the special file which contains the collections of typical irony. By comparing with our datasets, we will try our best to find out the 'false' emotion express and drop them from our datasets.

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.
- 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

   DONE

2. Perform sentiment analysis on cleaned (geolocated) tweets by March 26th

   DONE

3. Combine the datasets to produce a singular, larger set that has been cleaned by April 2rd

   DONE

Things Need to be Done in the Future:

4. Write code to get tweets by day by April 11th
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

Update so far:

First we are trying to collect all the data together. According to the dataset we found, we can analysis one of our datasets by using the EmotionLevel.

	EmotionLevel	TwitterID	PostTime	IsQuery	Username	Co
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	TheSpecialOne	@switchfoot http://twitpic.com/2y1zl - Aww
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. M
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its c
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at
5	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole
6	0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need
7	0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rain
8	0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope they didn't h
9	0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	@twittera que me mu
10	0	1467812416	Mon Apr 06 22:20:16 PDT 2009	NO_QUERY	erinx3leannexo	spring break in plain city... it's sno

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

In order to keep a neat environment to analysis data results, we have removed the unrelated attributes from the dataset and interrupting contents. There are three columns we want to pay highly attention to, which is EmotionLevel, PostTime and Content.

Out[23]:

	EmotionLevel	PostTime	Co
0	0	Mon Apr 06 22:19:45 PDT 2009	@switchfoot http://twitpic.com/2y1zl - Aww
1	0	Mon Apr 06 22:19:49 PDT 2009	is upset that he can't update his Facebook
2	0	Mon Apr 06 22:19:53 PDT 2009	@Kenichan I dived many times for the ball. I
3	0	Mon Apr 06 22:19:57 PDT 2009	my whole body feels itchy and like its (
4	0	Mon Apr 06 22:19:57 PDT 2009	@nationwideclass no, it's not behaving a
5	0	Mon Apr 06 22:20:00 PDT 2009	@Kwesidei not the whole
6	0	Mon Apr 06 22:20:03 PDT 2009	Need
7	0	Mon Apr 06 22:20:03 PDT 2009	@LOLTrish hey long time no see! Yes.. Rai
8	0	Mon Apr 06 22:20:05 PDT 2009	@Tatiana_K nope they didn't h
9	0	Mon Apr 06 22:20:09 PDT 2009	@twittera que me m
10	0	Mon Apr 06 22:20:16 PDT 2009	spring break in plain city... it's sn
11	0	Mon Apr 06 22:20:17 PDT 2009	I just re-pierced m
12	0	Mon Apr 06 22:20:19 PDT 2009	@caregiving I couldn't bear to watch it. A
13	0	Mon Apr 06 22:20:19 PDT 2009	@octolinz16 It it counts, idk why I did el
14	0	Mon Apr 06 22:20:20 PDT 2009	@smarrison i would've been the first, bu
15	0	Mon Apr 06 22:20:20 PDT 2009	@iamjazzyfizzle I wish I got to watch it v
16	0	Mon Apr 06 22:20:22 PDT 2009	Hollis' death scene will hurt me severely
17	0	Mon Apr 06 22:20:25 PDT 2009	about to file
18	0	Mon Apr 06 22:20:31 PDT 2009	@LettyA ahh ive always wanted to see ren
19	0	Mon Apr 06 22:20:34 PDT 2009	@FakerPattyPattz Oh dear. Were you drinkin
20	0	Mon Apr 06 22:20:37 PDT 2009	@alydesigns i was out most of the day so c
21	0	Mon Apr 06 22:20:38 PDT 2009	one of my friend called me, and asked to m
22	0	Mon Apr 06 22:20:40 PDT 2009	@angry_barista I baked you a cake but I
23	0	Mon Apr 06 22:20:40 PDT 2009	this week is not going as i had
24	0	Mon Apr 06 22:20:41 PDT 2009	blagh class at 8 tom
25	0	Mon Apr 06 22:20:44 PDT 2009	I hate when I have to call and wake peo

We separately count the tweets with 'high' EmotionLevel and 'low' EmotionLevel. This result will be the foundation of the analysis below. Also, it's our first step to analyze the data.

```
In [43]: print(count)

0    800000
4    248576
Name: EmotionLevel, dtype: int64
```

Then we compare the data with the datasets which contains the frequent positive and negative words in the Twitter. Collected the counts of the times that people use the positive words and negative words when they are in 'high' EmotionLevel and 'low' EmotionLevel.

```
In [79]: sum = 0
         for j in file1['Content']:
             for i in filepos.h:
                 if i in j:
                     sum+=1
         print(sum)

837799
```

According to the results we get in here, from 248576 twitters with 'high' EmotionLevel, there are only 837799 positive words have been use.

```
In [92]: sum = 0
         s=0
         for j in file1['Content']:
             for i in fileneg.h:
                 if i in j:
                     sum+=1
         print(sum)

983898
```

However, there are 903898 negative words have been used in the 'low' EmotionLevel, which showing that people are more likely to use several negative words in one twitter when they are in 'low' EmotionLevel. Different than our prospections, people use less positive words to express their emotions when they are in 'high' EmotionLevel. By selecting the sample of the 'high' EmotionLevel Twitter, we find out that people would rather use the indirectly expression, such like "The moon is so bright tonight!", instead of saying that "I love the moon tonight." Sometime it is hard to understand the deep meaning under people's twitters, if they choose to use indirectly way to express their emotions. However, we are going to build the sample file which contains the irony and typical exceptions of people's expression. By using this file, we will have a better and more accuracy data results in the future.

By far, we have done the data cleaning, tried to perform sentiment analysis on cleaned tweets and combine the dataset together. Although we met some problems like 'unreadable data file' when we try to use Pandas to clean the data. We have tried several solutions and finally figure it out in the end. Second roadblock is achieving the way to analysis the data. Even we have the plans before, we met some technical problems when we were trying to follow our plans. Since our datasets are all about tweets which contains many texts and emoji, it is different with other regular digit data analysis. After several discussions and coding, we figure out the relationship between words and emotion level. Besides that, we are going to find out the relationship between emoji and emotions. It will be the future work need to focus on.

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