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

ABSTRACT

With the increasing usage of internet, social media has promoted information dissemination in social network. By using sentiment analysis, there is a significant application in both industry and academic. It helps companies and media organizations improve user experience and promote economic development. This paper and research serves to provide a better understanding about human sentiment on social media, especially in Twitter. We investigate the utility of linguistic features for detecting the sentiment of Twitter message. Based on two data sets of more than 1.8 million tweets in total, we find that the sentiment expressed on Twitter changes by time, especially the emotional Twitter messages.

INTRODUCTION

Nowadays, social media have tons of users, and millions of people share opinions on different aspects of life everyday. This growing usage of social media increases the possibility of having new opportunities and improving business. By using information technologies to seek out and understand the opinions of others, the companies and media can track and analysis user's sentiment towards brands or products. Also, sentiments analyses can help users insights about how the public feels in regard to their business, products, or topics of interest. 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.

While there has been a fair amount of research and application on how sentiments are expressed in network, what the relationship is between time and sentiments has been less studied. In this paper, we begin investigating people's mood change according to the time period of a day. Sentiment is a feeling or emotion, an attitude or opinion. By using the relationship of time and sentiment,

1http://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.

companies can improve their ability of advertising and promote the information diffusion.

Also, we want to find the emotions behind a series of words and emoji, used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention. Based on the datasets, we do the research about how people use the Emojis or Emoticons to express their real sentiments, and what some downfalls in current language processing techniques regarding this are and what could improve them. This will help vader, the sentiment analysis tool we ended up using in the end, have a better preference on processing the sarcastic use of emojis and improve its score system.

Our project will serve as a study of sentiment analysis based on Twitter data. In this project, the research will draw motivation from people to understand and analyze the people's opinions more efficiently.

DATASET SET

We use several datasets with different Twitter messages 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. The features contain user name, IsQuery, tweet contents, post specific time, and sentiment level.

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

Table 1: Sentiment140 Dataset

- **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. This dataset has lots data about Emojis, which helps us detect how people use Emojis to express their real sentiments.

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

User Name	Nickname	Favs	Rts	Favs+RTs	Tweet content
MALUMA	maluma	2089	449	2538	Y DELE 💪 💪 ... Viernes por la nc
MALUMA	maluma	2006	425	2431	Hoy me dieron esta sorpresa en @Estacanon u
MALUMA	maluma	1847	426	2273	AGRADECER... Regla #1!! Gracias a los medios
MALUMA	maluma	1576	373	1949	😇😈 @ Mexico City, Mexico https://t.co/8I2
MALUMA	maluma	1550	313	1863	"MEDALLO PARAÍSO" nueva colección amelissa
Charlotte R	charlottero	853	499	1352	Do I look bloated? 😊 @amellywood & @
Tanya Burr	TanyaBurr	477	103	580	I'm snapchatting all day at Coachella so add m
MALUMA	maluma	356	70	426	Batman o SuperMan? @ Miami International A
Alfredo Flo	AlfredoFlor	243	153	396	Ya boy has arriveeddddd 😛 🤘 #coachella
Fran	_Franncis	237	100	337	Prom & dodger game tonight w/my lover
Tanya Burr	TanyaBurr	188	50	238	Morning at the pool before the Coachella mad
Chuck Com	chuckcome	172	51	223	I love being on tour and doing the post-game F

Table 2: Geolocated Twitter Dataset

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.

RELATED 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 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

MAIN TECHNIQUES APPLIED

- **DATA CLEAN/PREPROCESS**

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.
- There are some tweets have miscellaneous characters, which is an interrupting for our research and analysis. Therefore, we need to remove the data contains miscellaneous characters 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.
- Removing duplicate tweets from various data sources to keep the results accurate.
- 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 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.
- Convert fields to be easier to work with.

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

Out[23]:

	EmotionLevel	PostTime	Content
0	0	Mon Apr 06 22:19:45 PDT 2009	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	Mon Apr 06 22:19:49 PDT 2009	is upset that he can't update his Facebook by ...
2	0	Mon Apr 06 22:19:53 PDT 2009	@Kenichan I dived many times for the ball. Man...
3	0	Mon Apr 06 22:19:57 PDT 2009	my whole body feels itchy and like its on fire
4	0	Mon Apr 06 22:19:57 PDT 2009	@nationwideclass no, it's not behaving at all....
5	0	Mon Apr 06 22:20:00 PDT 2009	@Kwesidei not the whole crew
6	0	Mon Apr 06 22:20:03 PDT 2009	Need a hug
7	0	Mon Apr 06 22:20:03 PDT 2009	@LOLTrish hey long time no see! Yes.. Rains a...
8	0	Mon Apr 06 22:20:05 PDT 2009	@Tatiana_K nope they didn't have it
9	0	Mon Apr 06 22:20:09 PDT 2009	@twittera que me muera ?
10	0	Mon Apr 06 22:20:16 PDT 2009	spring break in plain city... it's snowing
11	0	Mon Apr 06 22:20:17 PDT 2009	I just re-pierced my ears
12	0	Mon Apr 06 22:20:19 PDT 2009	@caregiving I couldn't bear to watch it. And ...
13	0	Mon Apr 06 22:20:19 PDT 2009	@octolinz16 It it counts, idk why I did either...
14	0	Mon Apr 06 22:20:20 PDT 2009	@smarrison i would've been the first, but i di...
15	0	Mon Apr 06 22:20:20 PDT 2009	@iamjazzyfizzle I wish I got to watch it with ...
16	0	Mon Apr 06 22:20:22 PDT 2009	Hollis' death scene will hurt me severely to w...
17	0	Mon Apr 06 22:20:25 PDT 2009	about to file taxes
18	0	Mon Apr 06 22:20:31 PDT 2009	@LettyA ahh ive always wanted to see rent lov...
19	0	Mon Apr 06 22:20:34 PDT 2009	@FakerPattyPattz Oh dear. Were you drinking ou...
20	0	Mon Apr 06 22:20:37 PDT 2009	@alydesigns I was out most of the day so didn'...
21	0	Mon Apr 06 22:20:38 PDT 2009	one of my friend called me, and asked to meet ...
22	0	Mon Apr 06 22:20:40 PDT 2009	@angry_barista I baked you a cake but I ated it
23	0	Mon Apr 06 22:20:40 PDT 2009	this week is not going as i had hoped
24	0	Mon Apr 06 22:20:41 PDT 2009	blagh class at 8 tomorrow
25	0	Mon Apr 06 22:20:44 PDT 2009	I hate when I have to call and wake people up

Figure 1: Process of cleaning

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.

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
- Vectorizing tweet scores
- 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 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.

Classification and clustering

First we made SVM to classify tweets.

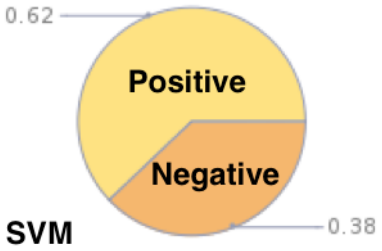

Figure 3: SVM Model

Classified tweets as positive, negative, or neutral along with subjective vs. objective using Vader (NLTK).

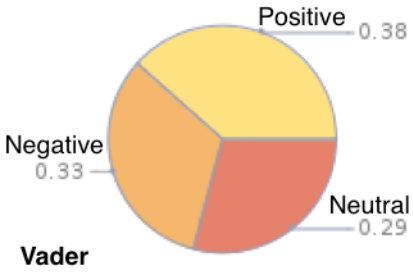

Figure 4: Vader Model

K-Means clustering with various 'K'.

TOOLS USED

We used 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.

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

- 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)

KEY RESULTS

The key result of our work is our analysis of the relationship between time and people's mood change. We were not much successful in finding the sarcastic use of emojis, but we compare the emojis with the preceding and following words in order to account for context.

Relationship between time and sentiment

From what we found, the sentiment expressed on Twitter does change throughout the day. For example, positive sentiment peaks in the morning while negative sentiment tends to remain relatively constant throughout the day. An interesting item we found is that in negative sentiment clusters in the morning, words like "shop", "purchase", and "buy" have a high frequency. However at night, those same words appear in positive sentiment clusters. Also, we found that tweets have a higher emotional intensity during mid-day times. We believe this is due to the fact that many people are at school or work and may be browsing online. While they are not interacting with friends as much, they are sharing a large amount of URLs that led us to believe people are looking at more news sources during these times.

The expression of Emojis in Sentiment

There are a lot of emojis present in tweets and they can be a bit ambiguous. Vader, the sentiment analysis tool we ended up using in the end, is a popular tool to process natural language from social media posts. We found that sarcastic use of emojis is still hard to interpret and Vader tended to incorrectly score these tweets. We also found a library within NLTK that can look at the preceding and following words in order to account for context.

MoreOver

Also, we did few interest research on positive and negative words usage of sentiment expression.

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)

903898
```

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.

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

APPLICATIONS

Due to what we found about peoples positive and negative sentiment, we could create models for advertising purposes to display ads for purchasing clothes for example in the evening when people seem to have a more receptive view towards spending money. Similarly, we could schedule the release of non-urgent news stories to target the largest amount of viewers.

For improving language processing, if we had more time we would've started this on our own. We believe that taking the set of tweets containing emojis and scoring them differently will produce more accurate results. Vader handles sarcasm already so given the context, it should be able to realize that there is ironic text surrounding the emoji so the value of the smiley face should add to the existing sentiment rather than contradicting it.

For positive and negative words using part, we did the research of sentiment with positive and negative words use. Also we discovered how these words express people's sentiment. Artificial intelligence and Machine Learning are the two of the most popular topics in computer science area and they both involved with language processing and need a language processing library. By researching positive and negative words and their sentiments, we knew satire and irony will totally change a word's attribute. By discovering the sarcasm use of positive and negative words, the artificial intelligence and language translating application will have a big step in approaching to humanoid.

REFERENCES

[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.

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

[4] Ley Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu, "Combining lexiconbased and learning-based methods for twitter sentiment analysis." Technical Report HPL-2011-89, HP, 21/06/2011.

[5] Ankit Kumar Soni, "Multi-lingual sentiment analysis of Twitter data by using the classification algorithms", Electrical Computer and

Communication Technologies (ICECCT) 2017 Second International Conference on, pp. 1-5, 2017.

[6] Stefan Stieglitz, Linh Dang-Xuan, "Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior," Pages 217-248, 08 Dec 2014.

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