Sentiment Analysis on Donald Trump

COMP90055 COMPUTING PROJECT(25 points credit)

due on 27/02/2019

(Group of 3, I am solo on summer course)

Dejun Xiang_349329

dxiang@student.unimelb.edu.au

Superviser: Pref. Richard O. Sinnott

DEJUN XIANG

I certify that:

- this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university;

and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.

- where necessary I have received clearance for this research from the University's Ethics Committee (Approval Number ....) and have submitted all required data to the School - the thesis is 2772 **words in length (excluding text in images, table, bibliographies and appendices).**

27/02/2019

Content:

Related Code:

https://github.com/Derek-Xiang/Project_Computing_Dejun_Xiang_349329

Dataset:

https://cloudstor.aarnet.edu.au/plus/s/qN1BgdMOG4hpuNl

Abstract

Nowadays, people's lives tend to be more and more digitised, which means more activities are conducted with cell phone converting the physical world to digital data. Sentiment analysis is a method using data collected from people to mining people's opinion, emotion and attitude towards some topic or item. Social media platform is a great place where people from all over the planet to demonstrate their thoughts and opinions on some specific issue. So two of the most active platform will be studies, which are 'Twitter' for English user and 'Weibo' for Chinese user. Since Twitter produces around 500 million of tweets daily which amounts to about 8TB of data [1], and Weibo has the ⅓ size of Twitter but with more users than the Twitter, which are quite valuable resources of the analysis mining sentiment on Donald Trump for this case.

Therefore, this report will explore how people think of Donald Trump by collecting data from these two platforms and, also, in order to make use of two platforms together, the comparison between English user and Chinese user's view on Donald Trump will be conducted, especially on the period before and after the outbreak of trade war between China and US. As a result that this is a group project of three, this report will be more focused on data retrieving and data pre-processing phase.

1.1 Introduction

Social media has gained tremendous attention in the past two decades, as the potential of helping people make decisions and sentiment analysis has been seen by more and more researchers and stakeholders like the politician.[2] In 2008, according to Cornfield, the great success of Barack Obama was attributed to the "online politics" strategy accomplished by his campaign team[3]. In addition to that, the problem of Indonesian Presidential Election in 2014 was also addressed by analysing social media[4], which before includes the human effort to collect data involving face-to-face interviews, phone calls and so on[6]. As results, the most popular platform in the world, Twitter, and the most user registered platform Weibo are used to collect natural text.

Sentiment analysis can be categorised into two approaches: Machine Learning based approach and lexicon-based approach[7]. Machine Learning based approach deals with this problem by using classification algorithms, which can be but not limited to SVM, logistic regression, K means etc, whereas lexicon-based approach uses sentiment dictionary with opinion words[8], and it already has every own matched polarity, so that the determination of one statement whether it is positive or negative can be calculated according to that dictionary straight away. This report will concentrate on the lexicon-based approach, and the tool used is TextBlob under the textblob package for English, and SnowNlp for Chinese.

There are mainly three schemes in lexicon-based approach: Manual Scheme, Dictionary based scheme, Corpus-based scheme. [7]

2. Dataset

Because the period around 23/03/2018 when the trade war happened is interested, so the raw data is all crawled from Twitter and Weibo from 1/1/2018 to 13/05/2018 on the keyword - Donald Trump. There are approximately 128 MB data collected from twitter.com, which has 482596 tweets, and the attributes include user name, tweet, number of likes, number of replies, number of retweets and dates.

Interm of weibo, in total, 362896 weibos are retrieved from weibo.com, which is about 197 MB, and it has features including user name, dates, number of likes, number of retweets, number of comments, comment, comment by who, and time of the comment.
The distinction between the two datasets is because of the different structure of the two websites.

Also, the corresponding dataset of these two after being pre-processed is included in this dataset. The file is in either csv or json form.

2.1 Sampling

Sampling is the process to select the sample from the large or streaming dataset[4], and it is classified as Uniform Sampling and Biased Sampling [9]. Uniform Sampling means the probability of selecting every item is the same, which will keep the distribution of the original

data, so it is what we need if we have to do sampling. Although this report does not concentrate on machine learning, we need a sampling algorithm to get samples to be labelled for preparing the next phase. The method we will use is Reservoir Sampler, the reference from Stream Computing:

```
 Reservoir algorithm
Let S[1..k] be an empty array
Let m be 0
For each item x
        ◦ Increment m {stream length}
        ◦ If m ≤ k ◦ Put x in S[m]
◦ Else
        ◦ Let r be chosen uniformly in [1..m]
        ◦ If r ≤ k, S[r] becomes x
When queried, output S
```

It will conduct a random uniformly selection, which is what we want.

4. Methodology

4.1 Data Retrieving

1) Weibo

The data is crawled mainly by using requests and BeautifulSoup. First of all, an agent pool is hired which is made of a bunch of 'PC identity', which is used to mimicry the header sent from our computer, rather than python 3.5, which is obviously a program, and each time when sending request, an agent that is randomly picked from the pool to build up the header. What's more, a logon cookie is used to mimicry a logon state, to make it look like human activity.

Secondly, by observing the structure of the searching URL, it is essential to figure out a way to build up the URL required so that we can access the data about the keyword on the specific period. What is to be cautious is that in between every two calls, there have to be 2 or 3 seconds waiting time, otherwise, the program will be easily detected as a robot, and the connection of the IP will be blocked.

After getting the source of every page, the extraction can be done by BeautifulSoup with find_all method to get all the tags, class or id to locate the information we need.

Finally, the data collected will be stored in MySQL local database with the structure we want.

2) Twitter

It is quite convenient and accessible to make use of Twython, which is the prime Python library which provides a natural means for accessing Twitter data.[1] It offers sufficient tools like User information, Twitter lists, Timelines in Twitter API, and also the User authenticated calls and application authenticated calls are included, as long as a development account has been registered. However, there is a limit. For past tweets, it only allows to get at most 5000 tweets, worse still, in the past one week. Although the streaming package is handy to get plenty of tweets, it just returns ongoing tweets which also has a limit, up to 1% of the whole traffic. In addition, all those data are not random uniformly selected, which means it is biased data, so this method was abandoned after some research done.

In order not to get IP blocked, selenium with webdriver is used to build the web spider. For this project, the chrome webdriver and firefox webdriver are used. The procedure is similar to the Weibo crawling. Firstly, create the URL on searching Donald Trump, and call the webdriver. It will open up a new window for the URL. Secondly, scrolling down to the bottom of this page, as twitter.com will not return the loaded contend until it is reached. This step is the reason why two browser weddriver is needed as if only using one for a long time, sometimes, the twitter may stop you from loading data on that page, which means you can get all but tweets. Next, after scrolling down to the bottom of the page, the find_elements_by method will be used to extract the information that is related. It could be by class name, XPath and so on. Finally, all the data collected, this time, will be stored in a dataframe, from pandas and numpy packages and then saved into a CSV file.

One trick that was used for the twitter crawling is to open multiple programs to crawl the data by various spiders as long as the internet speed is fast enough. It speeds up the process dramatically.

4. 2 Pre-processing

The raw data retrieved has lots of useless or even disturbing elements that have to be removed or changed until the left all is what is essential. Haddi et al. explored the role of text

pre-processing in movie reviews sentiment analysis through experiments that show the accuracy of sentiment classification may be significantly improved using appropriate feature and representation[5].

1) Remove Redundancy Data

Since the data is crawled by multiple processes, and the searching page of twitter sometimes will give back the tweets from the day before the day we are after, so data redundancy is quite a problem. This issue can be solved easily with pandas.dropredundancy.

2) Convert Data Type

The default data type is object, which is fine with string as pandas have str function, but in terms of other attributes like numbers and dates. It is essential to be converted to integer and DateTime data type so that the following calculation can be conducted. Not only the data type needs to be turned, but the format also has to be changed, as the number could be "1.5K" for 1500, or there is Chinese character ("日", day) standing inside the dates cell.

3) Remove Empty Cell

Facing the wild, raw data, it is undoubtedly familiar to meet this situation, which is the result of various causes such as data corrupted, webpage crawling error or merely the user did not fill it up, which happened in all the columns. For tweets and dates, it can be tackled merely by deleting. But for the number of likes, retweets and replies, the typical solution is to fill with mean or median. What's interesting about this dataset is that the way is more than ten times bigger than the median, which means some popular user attracts the majority of audiences, so the median fills them, to fit the distribution more lean to regular users.

4) Remove 'Reply Symbol + what it's at @'

In tweets or Weibo, there always can be found some statement like '@Donald Trump' or '#Donald Trump', which contributes nothing to the sentiment of this sentence, which will be noise in the following training process. So the regular expression is used to locate them and deleted. All this kind of remove step has to be followed by the step (3) to remove what is empty now, such as the figure.1 below:

@一只巨鲸　2018-01-06 02:13:49

@前列腺炎祖传秘方　2018-01-04 10:11:23

Fig.1

5) Clean the Tweet/Weibo

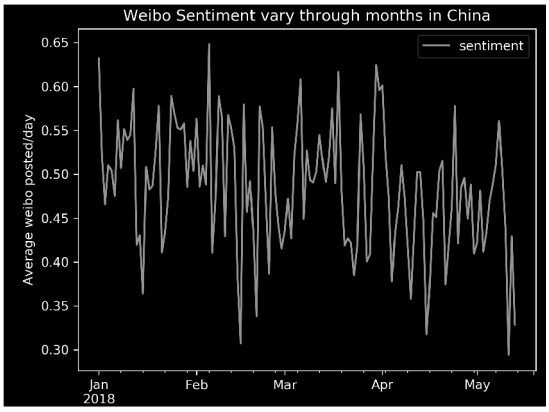
- For Weibo:
  - ❑ remove the punctuation other than Chinese character by re.
  - ❑ remove the stop words that are in the stopwords dictionary.
  - ❑ Splitting words in Chinese is a challenging problem, as different cutting may return distinct meaningful expression, but luckily we can tackle it with jieba, which is believed to be the best segmentor that we have run through, and it allows you to add words to its dictionary to adjust the context of our project.
  - ❑ Join up the cleaned words and return an array representing the cleaned weibo.
- For Twitter:
  - ❑ remove the punctuation other than English.
  - ❑ make all letters lower case
  - ❑ remove the stopwords by nltk.
  - ❑ splitting the word. This step is far easier than it with Chinese, as English is a segmented language. The spaces are naturally put between words.
  - ❑ combine the isolated words to return an array

4. 3. Class Classification

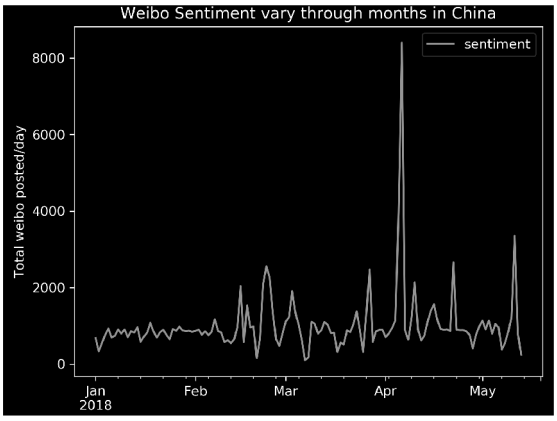
There is a tremendous way of classifying it is a positive or negative tweet as mentioned before. However, this report uses the TextBlob and SnowNLP to calculate the polarity of every tweet or weibo respectively. The category are 1 for positive, 0 for neutral and -1 for negative. The polarity will be added up together with the same date. So if the data is not biased, the result should be the sentiment on Donald Trump on that day overall. Also, the number of likes, the number of retweets and the number of replies can be the weight of that polarity, but the coefficient needs to be figured out by machine learning.

4. 4 Data to Graph

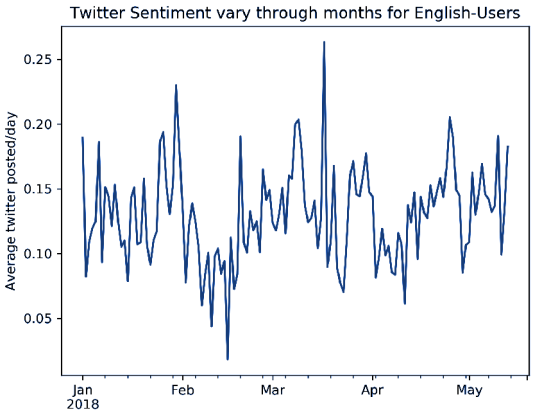
The axis of x is the month from Jan. to May in 2018, and the y axix represents either average sentiment of that day or the sum of that day, which is shown as below:



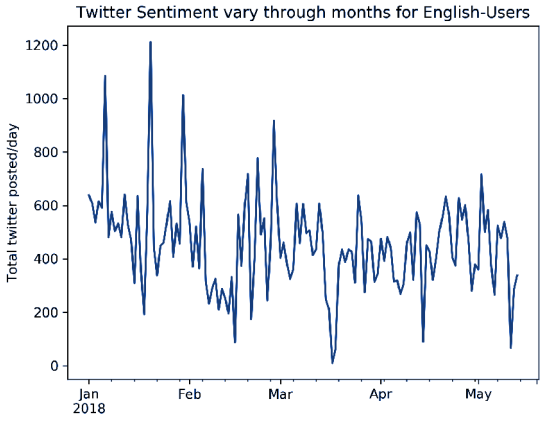
G.1 - Weibo - Mean



G.2 - Weibo - Sum



G.3 - Twitter - Mean



G.4 - Twitter - Sum

For Chinese user, who uses Weibo most, the break of the trade war on 23/03/2018 was not bother them a lot, but the trend of 'loving Trump' was stopped and started to going down after that. Also, at the head of Apr, talks on Trump increase dramatically. It is weird, and it can be believed the buzzer accounts turn up to try controlling the sentiment, which kind of did the job according to G1.

In terms of the Twitter user, although, it seems that the mean sentiment, which is from 0.05 to 0.3 every day is lower than Weibo which is from 0.25 to 0.65, the trend looks similar with time progressing. The time when it makes these two apart from each other is around late Mar., which is about the time of the trade war. From G3, it can be found there was a sharp

high sentiment climbing to the top at the affair breaking time, which can be said people quite like the idea of that. Moreover, according to G4, the prevailing trend actually goes down, which means English twitter user does not that concern about this issue as the news does. However, from G2, the Chinese user becoming more and more interested in this, which could because it indeed affect people's lives or the Propaganda department or the media start to move again.

Although the sum number may be different because of crawled data or some other side effect, the relative trend tells the story of what was happening.

5. What else can be done in this phase

5.1 Emoji Detection

When removing punctuation, the emoji is also deleted, which in fact presents attitudes or emotion and it could be an essential clue for classification. This could be done by finding an 'emoji dictionary' or a trained classifier to detect its corresponding polarity.

5.2 Buzzer Account Detection

Because the tweets or weibo are the voice from people, and also it returns back to people. We not only posting text on it, but we also gain information from it as well. So it provides the opportunity for those who intend to control or mislead people's mind. They hired human labour or robots to posting tweets that benefit them and twisted the fact. This will definitely affect the sentiment analysis. Ibrahim et al did the research on examining if Indonesian political Twitter messages reflect the real world activities of Indonesian Presidential Election [3].

It can be two categorized into two class, human hired and robots, whose accounts are merely created due to a special event [3]. It is also found in our project. As seen in figure 2 below:

182	乌拉拉的杨	你妈都是这么教你的?	2018-01-02 01:53:34
183	乌拉拉的杨	你美国爸爸啥时候争点气，把欠中国的钱还了。	2018-01-02 01:52:44
184	乌拉拉的杨	[眼泪]哪见过这么贱的畜生，中国劳动人民的血汗换几张擦屁股的印刷纸，还要欠着剥削者的恩么?	2018-01-02 01:52:02
185	乌拉拉的杨	美国没有网禁?	2018-01-02 01:49:06

Fig.2

These four messages are replies towards one post, which is written by one person, within 4 minutes. It is weird one person reply one post three times and all is negative, which is apparent that it is not a typical user. Since this is also a classification problem, we can have the following features to study with:

- The creation time of the account
- the frequency of posting during a specific period
- the number of followers/following
- the majority polarity of tweets and replies
- commonly used sentences(as the robots usually post text selected from a pool)

If we can also deal with these two problems, it is believed that the result will be more unbiased.

6. Conclusion

Overall, this report has thoroughly explored the process of how to conduct sentiment analysis on Donald Trump with Twitter and Weibo both, which includes data retrieving, data pre-process and some data mining with lexicon-based method, together with graphs. There are things that can be improved all the time. However, the prevailing trend can be found already, which does not need that precise at all. So, what we learned is that the scope of a project is essential, otherwise, many resources like computing force and time will be lost in vain.

Reference:

[1] Ahuja, S. and Dubey, G. (2017). *Clustering and Sentiment Analysis on Twitter Data*. 1st ed. 2017 2nd International Conference on Telecommunication and Networks: 2017 2nd International Conference on Telecommunication and Networks, pp.1-5.

[2] Bruno Ohana , Brendan Tierney," Sentiment Classification Of Reviews Using Sentiwordnet" 9th. IT&T Conference, Dublin Institute Of Technology, Dublin, Ireland, pp. 22-23, October 2009.

[3] Ibrahim, M., Abdillah, O., F.Wicaksono, A. and Adriani, M. (2015). *Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in A Twitter Nation*. 1st ed. Depok, Republic of Indonesia: 2015 IEEE 15th International Conference on Data Mining Workshops, pp.1348-1353.

[4] Priyanka1, Y. and Radha, S. (2019). *Sampling Techniques for Streaming Dataset using Sentiment Analysis*. 1st ed. Anna university-MIT campus: 2016 FIFTH INTERNATIONAL CONFERENCE ON RECENT TRENDS IN INFORMATION TECHNOLOGY.

[5] C¸eliktu˘ g, M. (2018). *Twitter Sentiment Analysis, 3-Way Classification: Positive, Negative or Neutral?*. Ankara, Turkey: 2018 IEEE International Conference on Big Data (Big Data), pp.2098-2102.

[6] A. Trihartono, "A vox populi reflector or public entertainer? mass media polling in contemporary indonesia," Procedia Environmental Sciences 17, pp. 928–937, 2013.

[7] Hailong, Zhang, Gan Wenyan, and Jiang Bo. "Machine learning and lexicon based methods for sentiment classification: A survey." Web Information System and Application Conference (WISA), 2014 11th. IEEE, 2014.

[8] Aung, Khin Zezawar, and Nyein Nyein Myo. "Sentiment analysis of students' comment using lexicon based approach." Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on. IEEE, 2017.

[9] Wenyu Hu, Baili Zhang, "Study Of Sampling Techniques And Algorithms In Data Stream Environments", 9th International Conference On Fuzzy Systems And Knowledge Discovery , pp. 1028 – 1034, May 2012.

[10] Nausheen, F. and Begum, S. (2018). *Sentiment Analysis to Predict Election Results Using Python*. 1st ed. Hyderabad, India: Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018), pp.1259-1262.