Analysis of Sentiment from Tweets Related to Manchester Football Team

Supervisor: Dr. Riza Batista Navarro
Daqian SHI   ID:10167702

1. Abstract

This report is written as a part of the requirements of Taster Project for MRes students. It is aimed to explore the rate of support for two football teams, Manchester United and Manchester City, by analysing the sentiment of tweets which mention them. The main programing language employed in Taster project is Python. JavaScript was also used to build a webpage showing the tweets on a geographic map.

2. Introduction

Social media has become a popular research area with the increasing popularity of social networking applications. They record huge amounts of personal information and also potential business-related information (Lei and Huan, 2010)[1]. Analyzing data from social media is becoming an important and general measure to learn more about a specific area or topic. Fire and Puzis (2012)[2] tried to analyze their employees' information from Facebook and LinkedIn, and found that the social connection of employees can be extracted as a network. Early research had predicted the stock market by analyzing the sentiment of tweets with an accuracy of 87.6% (Johan, Huina and Xioajun, 2011)[3].

The rate of support for a football team concerns football fans all over the world. With the beginning of the UEFA Champions League, football fans become more active on Twitter and publish tweets about their football teams. Can the rate of support for a team be detected by tweets? In this report, I investigate whether public tweets can be used to detect the sentiment of users, and furthermore, to analyze the rate of support for two football teams, Manchester United and Manchester City. I used a sentiment classification tool called TextBlob in Python to classify the sentiment of tweets from the 10th-17th November 2017 as Positive, Negative and Neutral. TextBlob is a collection of natural language processing tools like noun phrase extractor, classifier and sentiment analyser (Steven, 2017)[14] which can Satisfy all the needs for text processing in this project. A geolocation tool called Geopy was used to search for the geolocation of each tweet and mark them on the map. The results indicate that tweets with positive sentiments are concentrated in Manchester while Liverpool has the most twitter users with negative sentiments towards Manchester United

3. Tweets Collection and Pre-processing

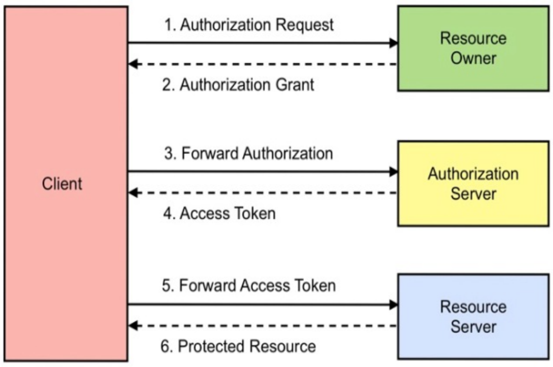

Fig. 3.1. Network communication between clients and servers

Tweets collection by REST API

The tweets were collected from twitter offline by using Python's Tweepy library, and were stored in a CSV file using the Pandas library with 14 main attributes like user name and text content.

To use Twitter's API (Kevin, 2009)[12] by Tweepy, several network communications would be required by Client to require the permission to use Twitter's API, as shown in Fig.3.1 (Twitter, Inc., 2017)[4]. Two possible APIs are available for collecting huge amounts of tweets: the REST API and the Streaming API (Twitter, Inc., 2017)[5]. In this project, I chose the REST API to collect tweets in a more efficient way since the Search method of the REST API can find historical tweets within the past week.

To simplify the sentiment analysis task, I only searched for tweets in English from all over the world. The keywords used in searching are team names, coach names and some of the players' names. A file was generated for each of Manchester United and Manchester City, containing 100,000 tweets.

An important attribute is the geocode, which will be used to map the tweets' distribution in Google Map. Geocodes are obtained by Geopy library which is a combination of common Map APIs in Python (Google Code, 2012)[15]. Five Map APIs like Google Map API and Nominatim Map API were run in parallel to search geocodes by attribute 'location' because of the rate limitation of each Map API is about 2500 per day, this method can improve the efficiency and accuracy of collecting geocodes and reduce Null value.

Pre-processing

Before sentiment analysis, the textual content of tweets was pre-processed and transformed into a standard format. Tokenisation is the most important and common step for text pre-processing whose purpose is to split sentences or text into single pieces (words or phrases) (Ronen, James, 2006)[6]. After tokenisation, remaining words were converted into lowercase form. Punctuations, emoticons and stop words like 'the', 'about' and 'myself' were also removed.

Frequent words and phrases

David (2002)[13] Argues that frequent words sequences can be used to confirm the main topic and the writing style of the text. To confirm that the tweets which I collected pertain to Manchester United (MU) and Manchester City (MC), I counted the frequent words and phrases within the two collections by using the Collections package. The result shows the most frequent word in the MU collection is 'manchester', accounting for 86,000 out of a total of 100,000 tweets. The former legendary coach Alex Ferguson and the current coach Jose Mourinho are mentioned 10,000 and 8,000 times respectively. The most frequent phrase in the MU collection is 'manchester united' (75,000). In the MC collection, the opponent team Arsenal was most mentioned (22,000) second to the home team name (80,000). The football star Sergio Aguero of MC appears 6,000 times. The most frequent phrase in the MC collection is 'manchester city' (70,000). Based on the frequencies of these words and phrases, the collected tweets have been found to be highly relevant to the main topic of the project.

4. Sentiment Analysis

Obtain sentiments based on lexicon

To obtain the sentiment of each tweet, I tried two possible methods. The first method is based on using the tool TextBlob to detect sentiments. The sentiment classifier underpinning this tool is based on natural language processing, in which a sentiment lexicon is employed to score every single tweet by detecting positive or negative expressions (Tetsuya and Jeonghee, 2003)[7]. The function for scoring is $S = \sum_{i=1}^{n} PW_i * DW_i * EW_i$ (PW: position weight, DW: degree word weight, EW: emotion word weight). If the score of a tweet is a positive value then this tweet is considered as having positive sentiment, otherwise, the tweet is considered as having negative sentiment. If the score of a tweet is 0, it means that the tweet contains neutral sentiment, e.g., containing only news. The sentiments of an evaluation set were marked manually and was used to test the accuracy of two methods. An average accuracy of 90% was obtained by the function: $P_a = \frac{TP+TN}{TP+TN+FP+FN}$ from confusion matrix (TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative).

Obtain sentiments based on machine learning algorithm

The second method I explored is based on training a machine learning-based sentiment classifier on the automatically generated results of TextBlob. In this case, a Naïve Bayes classifier was used, based on the Bayes theorem: $P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$ (Kevin, 2006)[8]. Upon extracting sentiments using the TextBlob sentiment classifier, the results are highly imbalanced – negative sentiment tweets only account for 10% in Manchester United set. N. V. Chawla (2002)[9] introduced a solution: ‘over sampling’ to balance all classes data by copying all the small size dataset. I drew inspiration from this work and performed over sampling. For this project, the training data of the Naïve Bayes classifier after over sampling trended to be average, which indicated that the accuracy loss from over sampling was ignorable. The final training set for this method comprised of a combination of 500 positive, 475 neutral and 482 negative tweets.

The accuracy of the Naïve Bayes classifier is 82% by the function: $P_a = \frac{TP+TN}{TP+TN+FP+FN}$, which is lower than the accuracy of the TextBlob although it also obtained useful statistics. The informative feature is a word or phrase that shows the correlation of the classification, while the rate of this correlation is called Positive and Negative Rate (PNR) which is the rate of the possibility of positive sentiment and negative sentiment respectively in this project. The sign of PNR presents the type of the sentiment and the figure of PNR presents the level of possibility of that sentiment.

By listing the most informative features of the Naïve Bayes classifier, the group of tweets which contain the word ‘go’ has the PNR of -331.6 as shown in Fig 4.1. It means that there is over 99% chance that any tweet relevant to Manchester United has positive sentiment if it contains this word. Meanwhile, the PNR of the phrase ‘to snub’ is -288 in chart of PNR statistic. Fig 4.2. shows the PNR of the most 10 informative words or phrases in the Manchester City collection. Some positive words like ‘unstoppable’ can be found in the positive words grouped with high PNR.

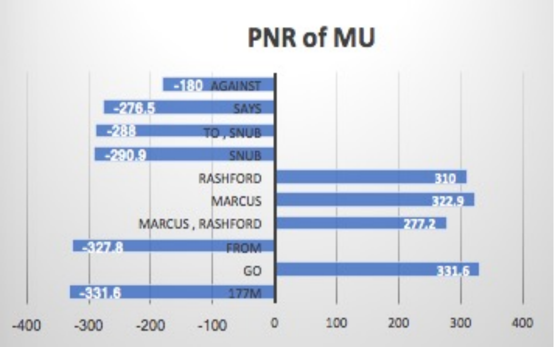

Fig. 4.1. Positive and Negative rate of informative words and phrases in MU

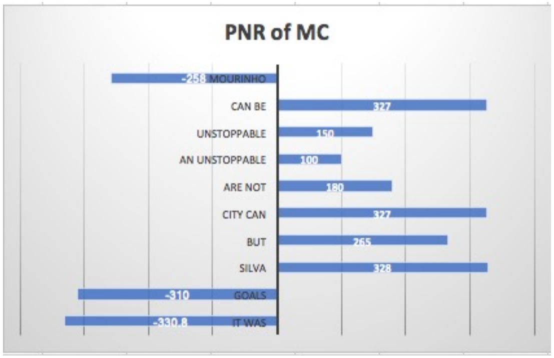

Fig. 4.2. Positive and Negative rate of informative words and phrases in MC

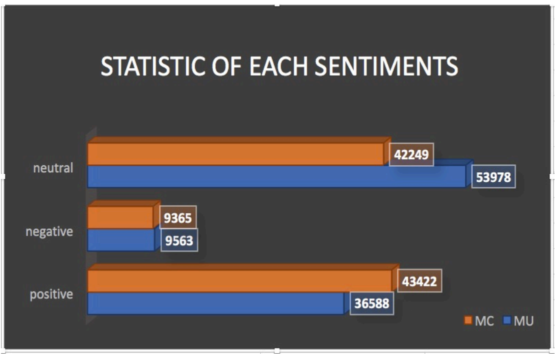

Fig. 4.3. Statistic of sentiment of each tweet set

Results

By comparing the accuracy of two methods, I decided to use the classifier based on the lexicon not only because of the high accuracy of extracting sentiments from tweets but also it is easier to implement, considering my time constraints. As a result, Fig. 4.3. shows the frequency sentiments within the MU and MC tweet collections. The trends observed from these two collections are similar, in which the number of negative tweets is the smallest and the number of neutral tweets is slightly higher than the number of positive tweets.

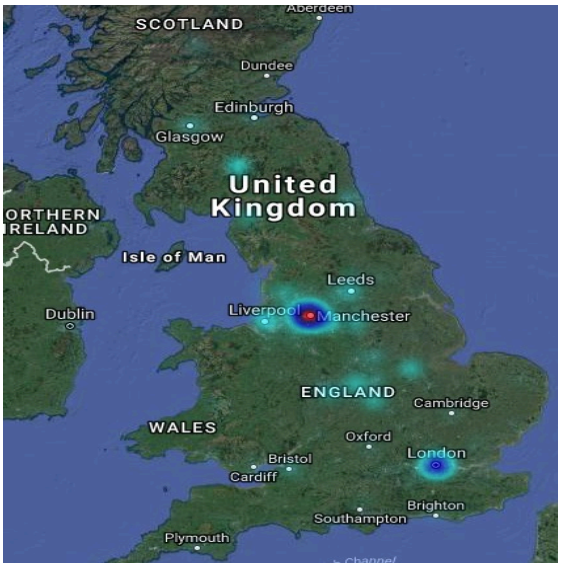

Fig. 4.4. Geolocation distribution of Positive tweets in United Kingdom

Fig. 4.4. shows a heat map (Ju et al., 2009)[10] visualising the frequency of positive-sentiment tweets which are relevant to Manchester United. It was built using the Google Map API (Yang and Wang, 2008)[11] which uses geolocation information to map all the points on global map. Manchester with surrounding areas and London have the most fans in the UK and positive tweets can be detected at the north of England like Glasgow. The only red colour in Heat Map is in the centre of Manchester, that means the most tweets about Manchester United from November 10, 2017 to November 17, 2017 are from Manchester, as one could expect

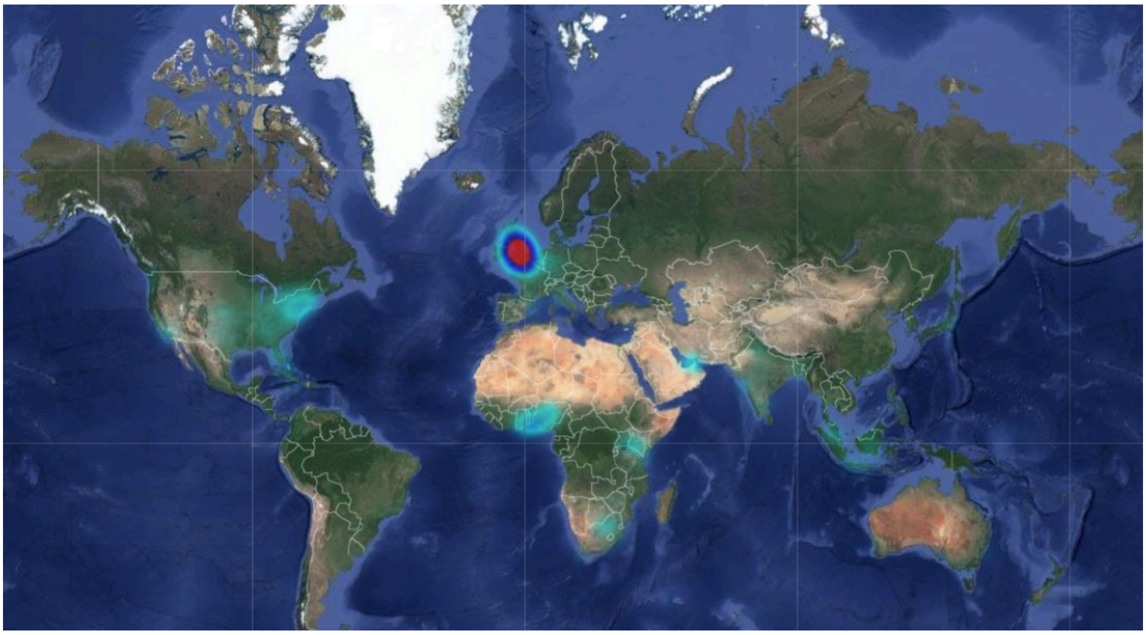

Fig. 4.5 Tweets distribution in global by Google Map

In Fig. 4.5., tweets with positive sentiment can be observed in several areas: the UK, the eastern and western coast of America, South Africa, Nigeria with surrounding areas, Kenya and United Arab Emirates. According to the most number of positive tweets in the UK, it has, as expected, the most number of MU fans. People in Asia, Eastern Europe and South America are not very active in posting tweets about these two football teams according to their few tweets. There is another interesting finding: the tweets with negative sentiment were also mapped using the Google Map API which showed that the most number of negative-sentiment tweets came from Liverpool. This makes sense as the battle between Liverpool and Manchester United has continued for decades and can be confirmed through analysis of tweets.

5.References:

[1] Lei Tang, &Huan Liu. (2010). Community Detection and Mining in Social Media. *Synthesis Lectures on Data Mining and Knowledge Discovery #3*.

[2] Fire, M., &Puzis, R. (2012). Organization mining using online social networks. *Networks and Spatial Economics*, page 1-34.

[3] Johan, Bollen., Huina, Mao., &Xioajun, Zeng. (2011). Twitter moods predicts the stock market. *Journal of Computational Science2(2011),* page 1-8.

[4] "Subscribe to your account activity". 2017 Twitter, Inc. [online]. Available: https://developer.twitter.com/en/docs/accounts-and-users/subscribe-account-activity/overview. [Accessed Nov 25, 2017].

[5] "Search Tweets". 2017 Twitter, Inc. [online]. Available: https://developer.twitter.com/en/docs/tweets/search/overview. [Accessed Nov 25, 2017].

[6] Ronen, Feldman., James, Sanger. (2006). *The Text Mining Handbook: Advanced Approaches in Analysing Unstructured Data,* page 60-61.

[7] Tetsuya, Nasukawa., Jeonghee, Yi. (2003). Sentiment analysis: capturing favorability using natural language processing. *Proceedings of the 2nd international conference on Knowledge capture*, page 70-77.

[8] Kevin P. Murphy. (2006). Naive Bayes classifiers, page 1-3.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer (2002) "SMOTE: Synthetic Minority Over-Sampling Technique", Volume 16, page 321-357

[10] Ju, Yeon Moon., Hyun, Jin Jung., Myeong, Hee Moon., Bong, Chul Chung., Man, Ho Choi. (2009). Heat-Map Visualization of Gas Chromatography-Mass Spectrometry Based Quantitative Signatures on Steroid Metabolism. *Journal of the American Society for Mass Spectrometry.*

[11] Yang, Tianliang., Wang, Liang. (2008). Telecommunical Base Station Information System Based on Google Map API. *Science Technology and Engineering,* page 208.

[12] Kevin, Makice. (2009). Twitter API: Up and Running, page 45-58.

[13] David L. Hoover. (2002). Frequent Word Sequences and Statistical Stylistics. *Literary and Linguistic Computing*, Volume 17, Issue 2, 1 June 2002, Pages 157–180.

[14] Steven Loria. (2017). "TextBlob: Simplified Text Processing" [online]. Available: https://textblob.readthedocs.io/en/dev/ . [Accessed Nov 25, 2017].

[15] "Geopy – A Geocoding Toolbox for Python – Google Project Hosting." (2012). Google Code [online]. Available: http://code.google.com/p/geopy/.