Software Design

10/2/2016

Jason Lee

Mini Project 1

Text Mining and Analysis

The data source that I used is Twitter. Because I am a huge fan of Manchester United, a professional football club in the Premier League, I wanted to know how people tweets about a recently signed player and the world's most expensive footballer Paul Pogba. The technique that I used is a sentiment analysis. By using such technique, I hoped to identity a sentiment polarity regarding his current performance in Manchester United.

To begin with, when I tried to use the sample program, or a code illustrated in the mini project to retrieve the tweets from the data source, I was not able to get the list of tweets. For me, the code was in the progress of running constantly, not showing the output. Therefore, from the pattern documentation, I used a code from Twitter search to get the data that I wanted. Instead of getting count of 10, I changed it to 30 so that I can get more list of tweets about Pogba. Before I ran the sentiment analysis, I had to pickle a data. Pickling data allows us to save the data and load back at a later point in time, rather than keep getting a new list of tweets. However, I had a trouble pickling a data when I did not store the 'tweet.text' somewhere outside of the for loop. Hence, I had to make a list after the for loop is over so that I can still access all the tweets and write them to a file.

After I got the list of tweets and pickled the data, I began the sentiment analysis. To use the sentiment analysis, I had to use the for loop so that I do not solely get the polarity of a single tweet. As I stored the tweets to pickle the data, I had to store the sentiment list. Inside of the second for loop, I appended 'tweet.text' to 'sentimentList'. Among multiple alternatives to analyze the sentiment polarity, I wanted to make the visualization much easier. In other words, I only took the average of the first part of the sentiment analysis. The reason why I disregarded the second part is that, to my delight, tweets are nearly all subjective. It is true that the average might not show the outlier of the sentiment analysis, yet it is much easier to visualize the result of polarity. Moreover, in the third box, I ran a sentiment analysis on a single tweet, "What is this? REF is right there and not even a yellow. The bias towards United/Pogba is unreal" to see how the results come out, which is (0.143, 0.268). It shows that this comment on Pogba is relatively neutral because the range of sentiment analysis is from -1 to 1.

Paul Pogba was a Manchester United player in 2011, yet got released to different team because he was not good enough. However, after five years, he returned and became the world's most expensive soccer player. Therefore, there are endless controversies about him whether he is worth that much or not, which I decided to run the sentiment analysis on him.

As the output below suggests, there are much of the negative polarities on Pogba and even most positive sentiment is 0.484. I initially believed that the result of this analysis would turn out to be very negative since he is currently not playing well in Manchester United. Nevertheless, my hypothesis

```
(0.14285714285714285, 0.26785714285714285)
(0.0, 0.0)
(0.039999999999999994, 0.49000000000000005)
(0.4, 0.8)
(0.16666666666666666, 0.25)
(0.4, 0.8)
(0.23472222222222222, 0.3944444444444445)
(0.0, 0.0)
(0.0, 0.0)
(0.0, 0.0)
(0.48409090909090907, 0.36363636363636365)
(0.2, 0.2)
(0.0, 0.0)
(-0.16666666666666666, 0.16666666666666666)
(0.0, 0.0)
(0.0875, 0.2375)
(0.14285714285714285, 0.26785714285714285)
(0.0, 0.0)
(0.0, 0.0)
(0.0875, 0.2375)
(0.0, 0.0)
(0.0875, 0.2375)
(0.14285714285714285, 0.26785714285714285)
(0.0, 0.0)
(0.0, 0.0)
(0.0, 0.0)
(0.0, 0.0)
(0.0, 0.0)
(0.14285714285714285, 0.26785714285714285)
(0.0, 0.0)
(0.2, 0.4)
```

was wrong. The first part of the sentiment analysis shows that they are pretty neutral. The average of sentiment polarity of 30 tweets about Pogba is 0.122.

As I have been working on this project, the result of the sentiment analysis did not come out as I expected. I have been watching Pogba playing since he returned to Manchester United and his performance did not show that he is the world's most expensive soccer player. There are still series of controversy regarding him. However, the average of sentiment analysis of 0.122 suggests that there are neutral comments about him. But, that does not mean his performance is in great shape because there are no single number that goes over 0.5. The neutral average of sentiment analysis does not clearly explain how public feels about his performance. It is true that 30 tweets that I stored cannot represent the performance of Pogba. However, before I pickled the list of tweets, I ran several sentiment analyses on him to get the big picture of the technique and the result followed: negative sentiments were predominant when I ran approximately 10 sentiment analyses. Nonetheless, the result I got for my project was very neutral. I would say that my project was not appropriately scoped to find out public's general thought on Pogba. Perhaps, if I get a chance to work on similar project in the future, I will change the count to 100 to get more variability. Additionally, it is difficult to know whether those who tweeted about him have sufficient and professional soccer knowledge. I would not simply conclude his performance based on the sentiment analysis of 30 tweets. Going forward, I am hoping to use what I learned from this project when Pogba gets the Man of the Match in the future. To be specific, I want to know whether the sentiment analyses on him become positive right after the excellent performance from him. Then, I can compare the result of the analysis for the beginning of the season, the time when he gets the Man of the Match in the future, and the end of the season.