North West Derby Twitter Analysis

2022 August match between Manchester United FC and Liverpool FC

Introduction

The event we selected for this analysis is the English Premier League match between Manchester United FC and Liverpool FC. We will try to use machine learning methods to analyse the tweets tagged on this match ranged between days before and after the game. Special interest is on how the Manchester United fans' perspectives or emotions on this game, and how they changed with the match result through the event.

Background

Manchester United FC and Liverpool FC are two of the most popular clubs in English Premier League. They both have won the top league trophy many times in the history. There always has been intense rivalry between the clubs. Thus, any match between these two teams is high-profile. The latest clash was round 3 for this season on 22 August 2022(London time). In recent years, Liverpool has had strong performance while Manchester United has suffered their poorest, especially they lost the first two rounds consecutively. Anger, frustration and negativity was among the United community before this game, and United was not expected to be the better team. However, the United won this game surprisingly by 2-1.

Data Collection

In this analysis, we use REST API on twitter instead of streaming approach. There are two main reasons for this. One is that our study just needs a relatively small scale (thousands) of tweets on this event. The other is that this is a recent event on 22 August 2022. The REST API can give us last 7 days tweets.

Query

After the API registration and set up, we use search_recent_tweets to extract the data. The query we use for our search is set to English language only, and must include both "#munliv"(the match event) and "#mufc" (Manchester United), as we try to focus more on the Manchester United fans or supporters. Also, we exclude retweets and tweets containing links to avoid collecting tweets for commercial advertising or online streaming or ticket selling, as well as scaling down the number of tweets we are going to collect, since this study is focusing on text.

Collections and initial processing

We firstly ran the query on 24 August 2022. The run gave us 6059 tweets dated from 17 Aug to 24 Aug. The second run of query was on 27 Aug to have another 60 tweets on 24 and 25 August. Then we bind these two data frames to one as our raw data and remove some duplicates on 24 August from two extractions. The raw data has 6,092 unique tweets. From *table 1*, we can see that most of the tweets were on match day 22 August.

Since we only captured 16 tweets partially on the day of 17 August due to the limitation of 7 days, we exclude these 16 tweets. Plus, we want to focus as much as possible on Manchester United only. Thus,

Table 1: Raw data tweets by date

	Counts of tweets
2022-08-17	16
2022-08-18	69
2022-08-19	141
2022-08-20	42
2022-08-21	136
2022-08-22	5261
2022-08-23	367
2022-08-24	45
2022-08-25	15

we remove the tweets tagging Liverpool FC at the same time. After the removal, we have 5952 tweets for our analysis.

Raw data is processed as pandas data frame with two columns – created_at and text. We also change the time zone to Europe/London to align with the match time zone. This will make it easier for us to extract the tweets on event time as well. Furthermore, we add date and time columns by separated the datetime formatted created_at column. Raw data is saved as a csv file.

Pre-processing and Data Cleaning

To process the text data, we will apply the following steps to each tweet content.

- Set all the letter to lower case.
- Tokenize the text using TweetTokenizer from nltk package.
  This is the process to separate sentences into word by word by space or punctuations. The advantage of using TweetTokenizer is that it is a special tool which can pick up the "tweet language".
- Remove leading and ending whitespaces for each token.
- Remove stop words based on nltk package.
  Stop words are frequently used words but having less important in meanings. For example, "is", "the" or "and", this category of words can be removed without impacting the understanding of the context.
- Remove any punctuations or special strings like 'rt', 'via', '...', '…', '"', '"', '`'.
  Special strings can be used frequently and popularly, but without actual meaning to the text. If we do not remove them, they would be appearing in our high frequency list.
- Remove any digits or fractions or hyperlinks.
  Numbers are not considered in this analysis. Although we have had excluded links in our tweets retrieval process. We will still remove any strings with pattern of "http".

Analysis Approach

To understand how the Manchester United fans' focus and thinking on this game through the days and how they react during the match time, we will apply two main analysis methods. One is sentiment analysis, and the other is topic modelling.

Sentiment Analysis

We will use VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicons for sentiment analysis. VADER can well understand texts containing emoticons, slangs on social media settings and does not require any training data. Steps are:

Preparation

- Prepare VADER analyzer for nltk package and download the up-to-date VADER lexicons.
- Load our data
- Each tweet text goes through the pre-processing and cleaning process.
- Apply polarity scores method on each processed text and then get the 'compound' score
- Add a score column in our data frame to store the compound score for each tweet
- Categorise the compound score to positive (score > 0.05), neutral (score between -0.05 and 0.05) and negative (score < -0.05)

Date trend analysis

- Create a pivot table reflecting counts of positive, neutral and negative on each date

- Calculate each date's negative rate (negative counts on total), positive rate (positive counts on total) and negative to positive ratio by using negative count divided by positive count.
- Plot the rates and the ratio and interpret the results

Match real time analysis

- Our real time analysis will use the subset of tweets during the time between 20:00 and 22:00 22 August 2022 when the match was played. 3200 tweets were posted during those 2 hours.
- Categorise the time into 15 minutes blocks from 20:00, e.g. 15 meaning tweets between 20:00 and 20:15, 30 meaning tweets between 20:16 and 20:30, and etc.
- Calculate each 15-minute block's negative rate (negative counts on total), positive rate (positive counts on total) and negative to positive ratio by using negative count divided by positive count.
- Plot the rates and the ratio and interpret the results

Topic modelling

We will use Latent Dirichlet Allocation(LDA) model to do the topic analysis. One advantage of using LDA here is that it does not require us to provide topics and it can detect topics. All we need to do is setting how many topics we would consider. LDA also performs very well with tweets. Steps are:

- Load our data
- Each tweet text goes through the pre-processing and cleaning process and save into a list.
- Set number of topics to 6, considering we are event focused analysis as well as the context is a specific type of sport event.
- Set number of words/features to 1000.
- We will use Term Frequency–Inverse Document Frequency (TF-IDF) vectorizer to transform our data. The main reason is to reduce the overlapping on topics comparing to TF only method.
- Display and visualise the topics

Analysis & Insights

Sentiment Analysis

As mentioned before, we will firstly investigate what the trend is from 18 August, which is 4 days before the match, to 25 August which is 3 days after the match.

Date trend analysis

a. Overview

From *table 2*, we can see overall fans posted tweets on the match day. 23 August has the second largest number since United won and the excitement the fans would have from the game. But the attention or excitement did not last long. It dropped significantly two days after the game. Possible reason for that is the game was played on night of Monday and United has the next game on Saturday noon. The focus could be moved to the upcoming game quickly.

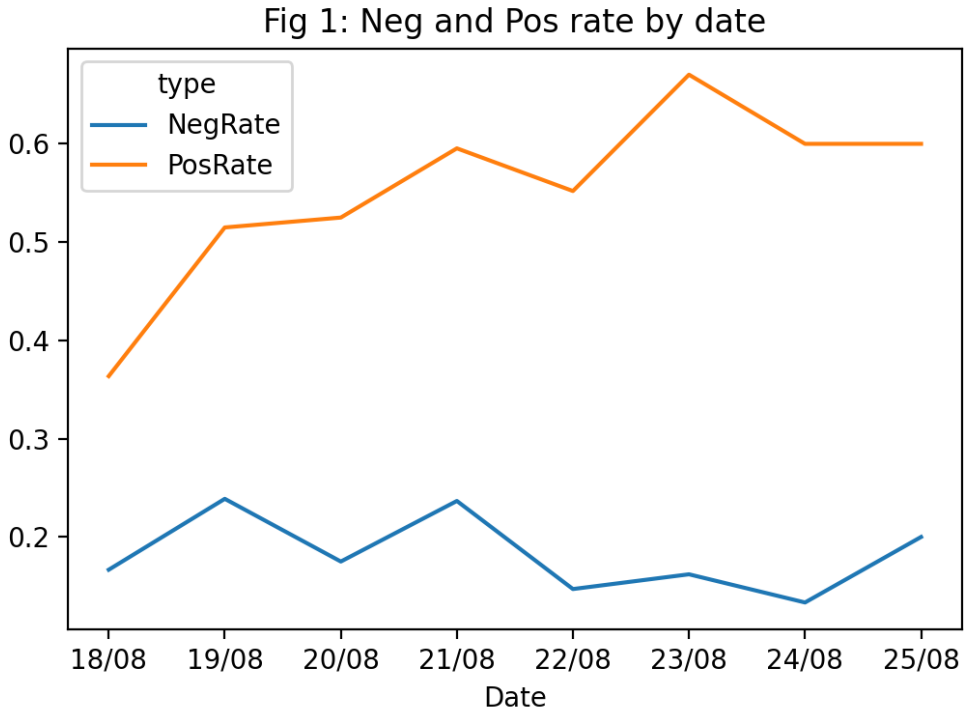
Given we try to focus on Manchester United fans or supporters only (by using #mufc and excluding #liverpool), we can expect more positivity on their club. Generally, this game added positivity to the United fans if we look at the numbers of the rates and ratio. We will go to details next section.

Table 2: Daily tweet count by date and type

Date	Negative	Neutral	Positive	total	NegRate	PosRate	NegPosRatio
18/08	11	31	24	66	0.166667	0.363636	0.458333
19/08	32	33	69	134	0.238806	0.514925	0.463768
20/08	7	12	21	40	0.175000	0.525000	0.333333
21/08	31	22	78	131	0.236641	0.595420	0.397436
22/08	759	1554	2850	5163	0.147008	0.552005	0.266316
23/08	58	60	240	358	0.162011	0.670391	0.241667
24/08	6	12	27	45	0.133333	0.600000	0.222222
25/08	3	3	9	15	0.200000	0.600000	0.333333

b. Negative and Positive proportions

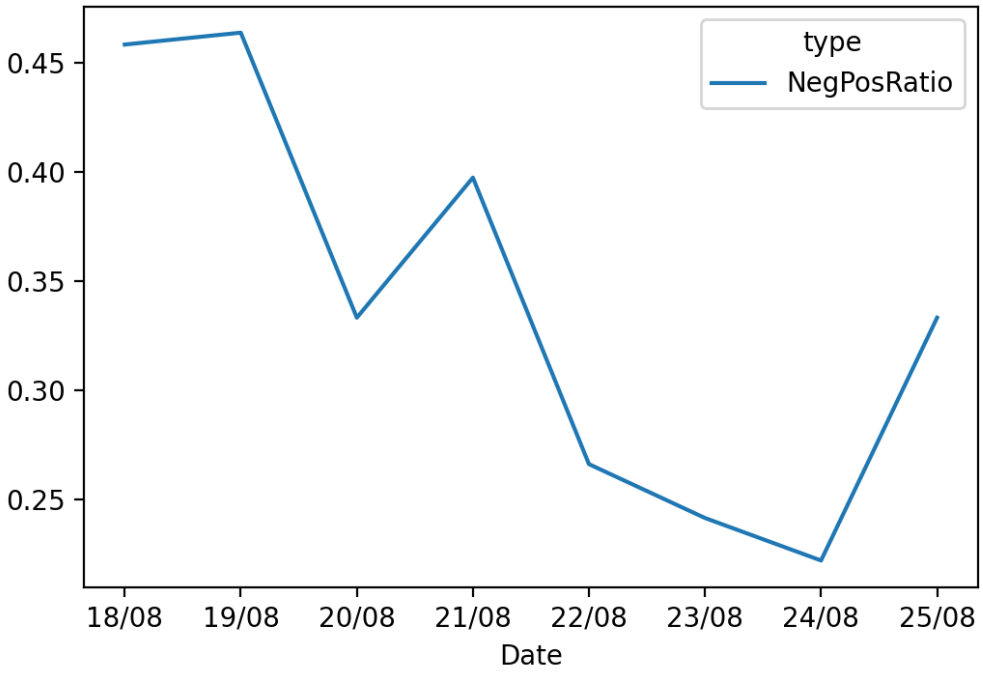
From *Fig 1*, we can see positive rate was quite low on 18 August following the shocking loss to Brentford on 14 August. Both positive and negative trends changed from the match day of 22 August. On 23 August, the positive rate reaches its peak. And from 25 August, the emotions of both sides seem to be resumed to the levels before the game.



c. Negative to positive ratio

More clearly, this method shows about the negativity reduction due to the match. From *Fig2*, we can see that pessimism was relatively high on 18 and 19 August. And between 22 and 24 August, the ratio shows a very clear declined trend due to the win to Liverpool. From 25 August, it has rebounded to the level before the game.

Fig 2: Neg/Pos ratio by date



Match real time analysis

a. Overview

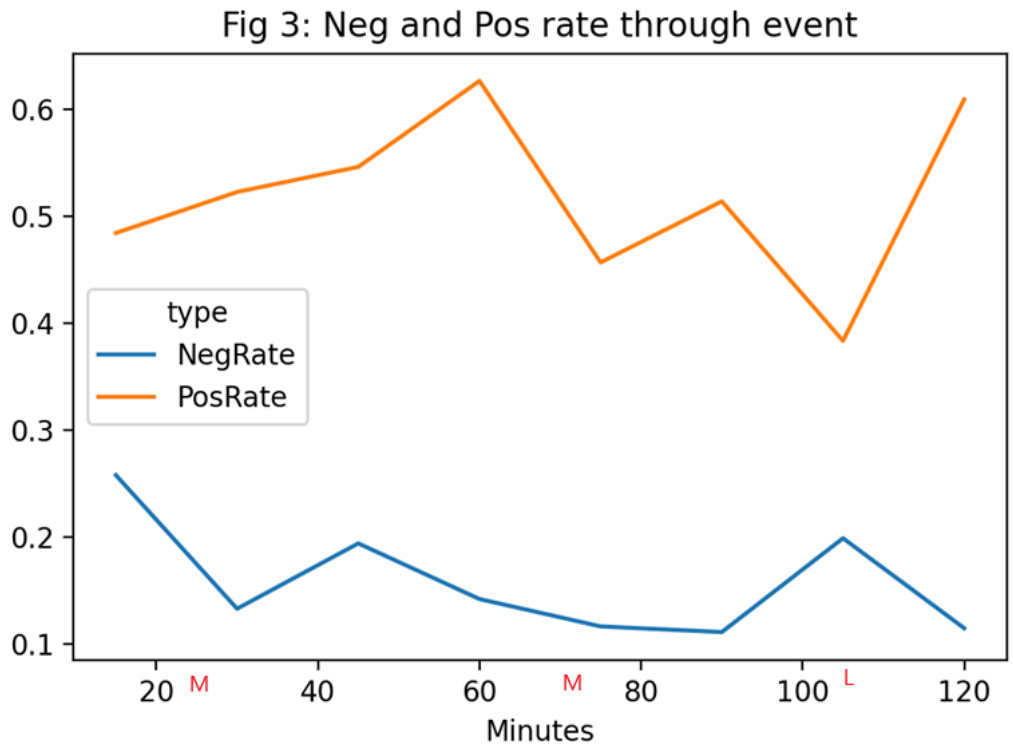
The match started at 20:00 and finished before 22:00. There are 3,200 tweets posted during this time. The timeline of the game is Manchester United scored at 16th and 53rd minutes of the match while Liverpool scored at 81st minutes. As we categorised them into 15-minute blocks, block 30 and block 75 are when Manchester United scored while block 105 is when Liverpool scored. We can see the numbers of tweets boosted whenever Manchester United scored. (*Table 3*) Tweets are more on half time break or right after full time which makes sense since fans did not need to watch to game.

Table 3: Daily tweet count by date and type

Minutes	Negative	Neutral	Positive	total	NegRate	PosRate	NegPosRatio	
15	67	67	126	260	0.257692	0.484615	0.531746	
30	72	187	284	543	0.132597	0.523020	0.253521	United goal
45	50	67	141	258	0.193798	0.546512	0.354610	
60	68	111	301	480	0.141667	0.627083	0.225914	Half time
75	50	184	197	431	0.116009	0.457077	0.253807	United goal
90	31	105	144	280	0.110714	0.514286	0.215278	
105	58	122	112	292	0.198630	0.383562	0.517857	Liverpool goal
120	75	181	400	656	0.114329	0.609756	0.187500	Full time

b. Negative and Positive proportions

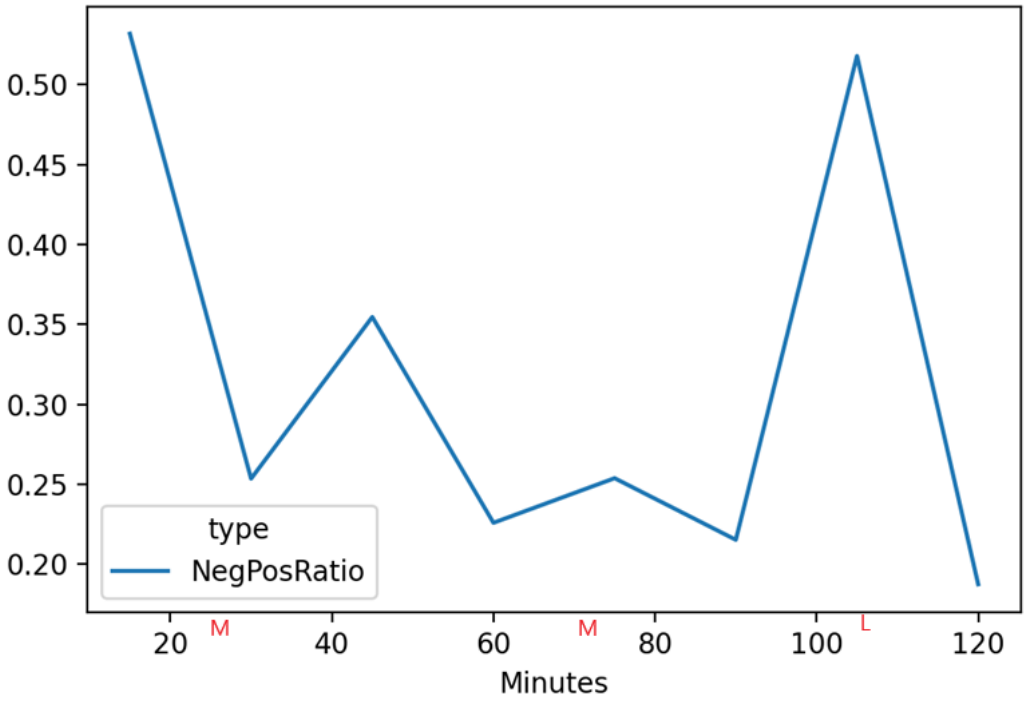
From Fig 3, the goals are marked as M or L for United or Liverpool side on the time blocks. We can clearly see the trends that United goals raised the positive rate and suppress the negative rate while the Liverpool goal did drag down much of the optimisation of the United fans. Positive peaks at half time break and right after full time. That means more fans are reflecting the first half and whole match excitement and happiness for the lead or the win of the game.



c. Negative to positive ratio

Fig 4 shows more fluctuation on the negative to positive ratio through the event and how it related to the goal scoring for either side. Two United goals pull down the ratio. However, the biggest observation here is the ratio dropped from more than .5, the highest, at the start of the match, to under .2, the lowest, at the end of the match. This reflects the result really cheered up the United fans. For that, we might also suggest that United fans did not expect the win. To add to this suggestion, we can see how quickly the ratio went up to close to the highest when Liverpool scored one goal while actually United still led the game and only 10 minutes left for the match.

Fig 4: Neg/Pos ratio through event



Topic modelling

As mentioned before, we use LDA model to detect top 6 topics. All the 5952 tweets are all processed together.

Results

From Fig 5, we can see the model picked up this game (topic 2) and the current situation that United captain Harry Maguire is struggling with a place in the team (topic 0). Also, the new signing of Casemiro has been detected (topic 4). The fans' long unhappiness on club owner the Glazers is a topic that the model would not miss (topic 5). These are in fact the current heat topics of Manchester United. Next, we will see some visualisations.

Fig 5: Top 6 topics with 10 most frequent words for each

```
Topic 0:
maguire bruno shaw captain starting manutd rt getting harry varane
Topic 1:
martinez sancho rashford martial malacia lisandro varane elanga goal salah
Topic 2:
united liverpool glazersout lfc manchester old trafford win glazers man
Topic 3:
want half bench ronaldo maguire second glory united fight way
Topic 4:
casemiro pl mctominay line fred gea mnf god watching ball
Topic 5:
glazersout manutd match game good united ggmu man emptyoldtrafford tonight
```

Visualisations

From Fig 6, we can see that there is no overlapping for our top topics, thanks to the TF-IDF method. The topics are well separated and represents a group of words.

Fig 6: Top 6 topics intertopic distance and most teams

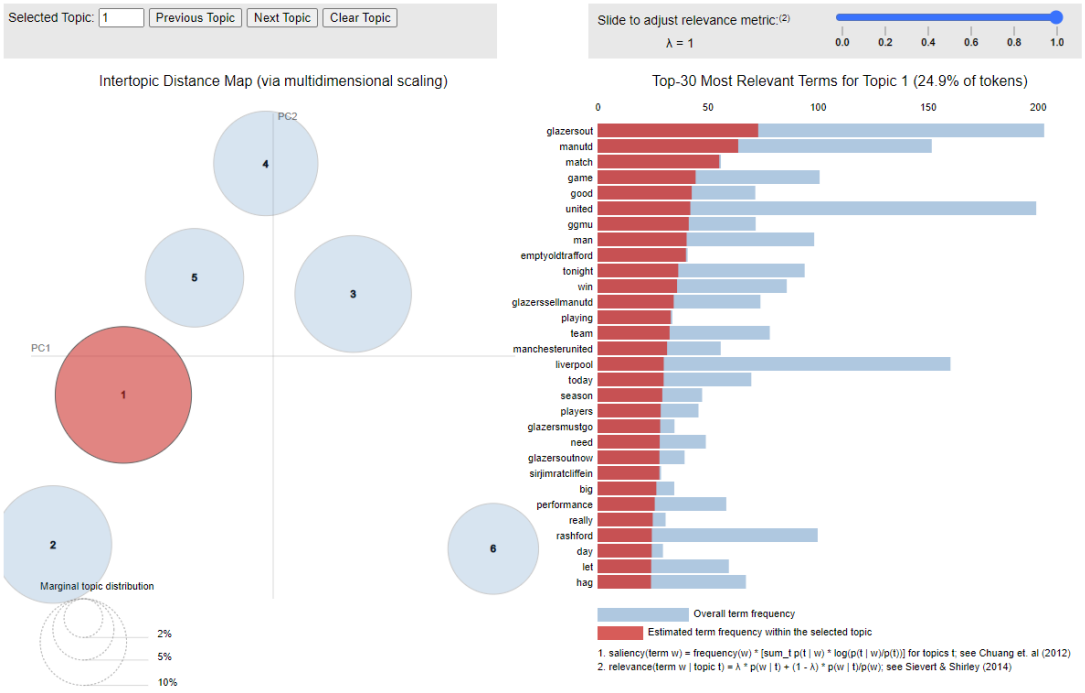


Fig 7 shows us the word clouds of these 6 topics. These visuals give us more direct descriptions of each topic, like, Maguire, player names, game between Liverpool and United, Casemiro and "Glazersout".

Fig 7: Top 6 topics word clouds

Conclusion

In this study, we have used sentiment analysis and topic modelling to investigate what Manchester United fans were focusing and how their emotions changed before, during and after the match against Liverpool. Both methods work well on understanding these tweets. The sentiment analysis showed us how the emotion trending along with the progress of the game. It really confirms that football is really such an exiting sport that can definitely affect people's emotions. The topic modelling successfully gave us the heat topics Manchester United fans are caring about.

Limitations in this analysis include a. the query we used to identify United fans is based on the assumption that Manchester United fans will tag MUFC but not tag Liverpool at all; b. there are still some phrases which were not identified correctly. For example, "A big performance for" should be a positive comment while it was deemed as neutral.

Further study could be conducted. One is to improve the sentiment model performance, the other is doing sentiment analysis on each topic.

References

[1] Shashank Kapadia, "Topic Modeling in Python: Latent Dirichlet Allocation (LDA)", medium.com, 2019. [Online]. Available: https://medium.com/towards-data-science/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0. [Accessed: 21- Aug- 2022].

[2] Alan Jones, "Sentiment Analysis of Tweets", medium.com, 2020. [Online]. Available: https://medium.com/towards-data-science/sentiment-analysis-of-tweets-167d040f0583. [Accessed: 21- Aug- 2022].