Arteen Zahiri, Spencer Griffin
Professor Zhi Li
MIS3640-01
31 October 2019

Assignment #2 Project Writeup and Reflection

## 1. Project Overview

Our main objective was to create a sentiment analysis of Piers Morgan's tweets relating to Arsenal. Piers Morgan is a popular British broadcaster and journalist known for his outspoken views on politics and his favorite football club: The Arsenal Gunners. We are also staunch Arsenal supporters and we personally follow Piers Morgan on Twitter, making this a project of personal relevance for both of us. We were particularly interested in finding trends relating to the timing of his tweets as well as the polarity of his tweets when he mentions Arsenal. Not only is Piers extremely active on Twitter, but he is a passionate and emotional fan, making him a perfect candidate for such an analysis.

## 2. Implementation

### *2.1. Web Scraping*

The first major component of our code was to retrieve the data from Twitter's API. We decided to use Tweepy for our twitter library after conducting research. According to the python online community, Tweepy is easier to use compared to Twython. There is also more information available which helped us learn how to use it. We wrote down our keys in a separate python file named credentials.py. We imported everything from the credentials.py file into our piersmorgan.py file. The next step of fetching data was to set up our API. In a file called get_tweets.py, we used our imported access keys to authenticate and return the Tweepy API. Using a tweepy extractor, we fetched and returned all tweets from Piers Morgan (limited to his latest 3210 tweets). In order to save us time when running the file, we stored the tweets in a list, pickled them, and dumped them into a separate pickle file (tweets_by_piersmorgan.p) that we could use in piersmorgan.py.

### *2.2 Data Pre-Processing*

The second major part of our code was to process the tweets based off of our keywords ('Arsenal', 'Emery', 'Xhaka', 'Ozil') and create a dataframe to contain the tweets and vital information about them. Arsenal is the name of the team, Unai Emery is the head coach, and Granit Xhaka and Mesut Ozil are two of the club's most notable players whom Piers frequently tweets about. The first step in this process was to load the pickled data we had retrieved previously. Then, we processed tweets using our list of keywords relating to Arsenal. We created a new list of tweets containing these keywords which would be later used for our analysis and dataframe. We then created functions in order to retrieve metrics such as the number of likes, number of retweets, the date the tweet was created, the source from which it was created, and more. We sorted these metrics into a dataframe using the pandas module and created columns using numpy.

### *2.3 Sentiment Analysis & Visualization*

The third and final component of our code involved the analysis and visualization of our processed tweets. Most of our output originates from this code. Since likes and retweets are the foundation for Twitter's algorithm, we decided to display the most liked and retweeted Piers Morgan tweet relating to Arsenal. Furthermore, we included our sentiment analysis in this portion of the code. We used textblob to score the processed tweets based on polarity. To maximize accuracy, we created a function that cleans the tweet texts by removing links and special characters before processing the tweets. Once this was completed, the polarity scores were added as a separate column to the dataframe (SA). With the new column, we printed out the most recent 10 tweets from the dataframe along with the metrics. The percentage of negative, positive, and neutral tweets was also on display as we wanted to observe the distribution of sentiment. Finally, using pandas and matplotlib, we created two time series visualizations. The first displayed the length of tweets over time while the second displayed the tweets and likes superimposed onto each other.

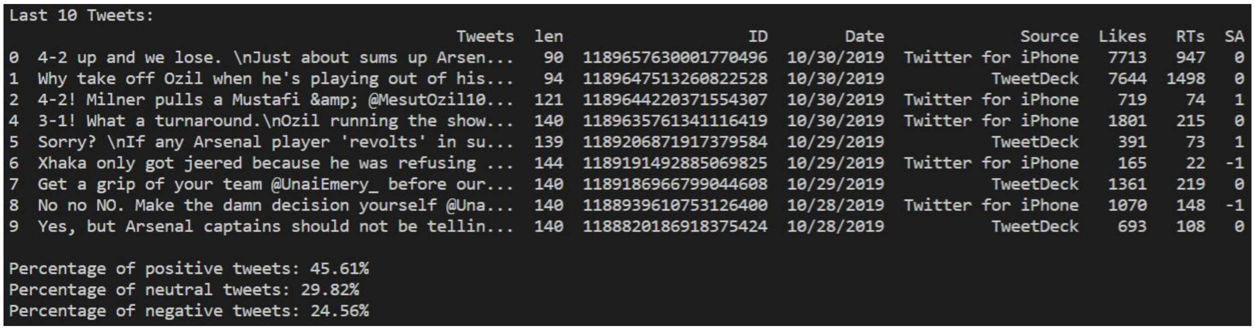
## 3. Results & Key Insights

### *3.1 Text Analysis Insights*

We learned that Piers Morgan gets more engagement (likes and RTs) on his tweets when he mentions controversial figures and uses inappropriate language. His most liked tweet (~30k likes) mentions Megan Rapinoe, a controversial women's soccer player who is an advocate for equal pay. His most retweeted tweet (~3.3k RTs) is him slandering a player, Granit Xhaka, for his poor performance. Piers even goes as far as using the 'f-word' in his tweet. While we do not condone this language, we certainly do not blame him for disliking Xhaka, who has been an embarrassment this season. The percentage distribution for polarity is rather intriguing. Piers' tweets are broken down as approximately 46% positive, 30% neutral, and 25% negative. Since both of us follow him, we find this highly skeptical. In our experience, we believe that most of Piers

Morgan’s tweets are negative in nature. The reason being that he drives more engagement when he feeds off the negative emotions of his followers. Thus, we believe that there may be a flaw in the text blob sentiment analysis.

Tweets Analysis:



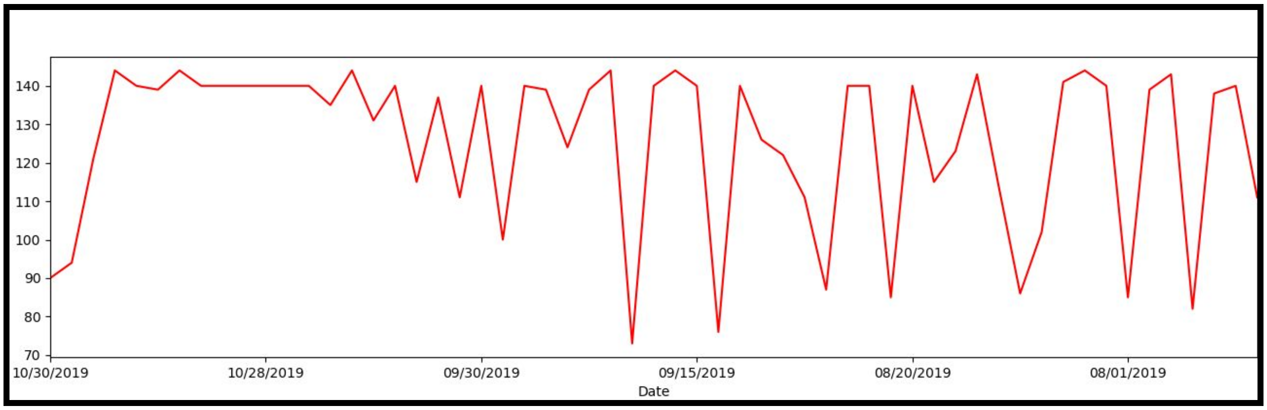
```
Last 10 Tweets:
                                              Tweets  len                   ID        Date              Source  Likes   RTs  SA
0  4-2 up and we lose. \nJust about sums up Arsen...   90  1189657630001770496  10/30/2019  Twitter for iPhone   7713   947   0
1  Why take off Ozil when he's playing out of his...   94  1189647513260822528  10/30/2019           TweetDeck   7644  1498   0
2  4-2! Milner pulls a Mustafi &amp; @MesutOzil10...  121  1189644220371554307  10/30/2019  Twitter for iPhone    719    74   1
4  3-1! What a turnaround.\nOzil running the show...  140  1189635761341116419  10/30/2019  Twitter for iPhone   1801   215   0
5  Sorry? \nIf any Arsenal player 'revolts' in su...  139  1189206871917379584  10/29/2019           TweetDeck    391    73   1
6  Xhaka only got jeered because he was refusing ...  144  1189191492885069825  10/29/2019  Twitter for iPhone    165    22  -1
7  Get a grip of your team @UnaiEmery_ before our...  140  1189186966799044608  10/29/2019           TweetDeck   1361   219   0
8  No no NO. Make the damn decision yourself @Una...  140  1188939610753126400  10/28/2019  Twitter for iPhone   1070   148  -1
9  Yes, but Arsenal captains should not be tellin...  140  1188820186918375424  10/28/2019           TweetDeck    693   108   0

Percentage of positive tweets: 45.61%
Percentage of neutral tweets: 29.82%
Percentage of negative tweets: 24.56%
```

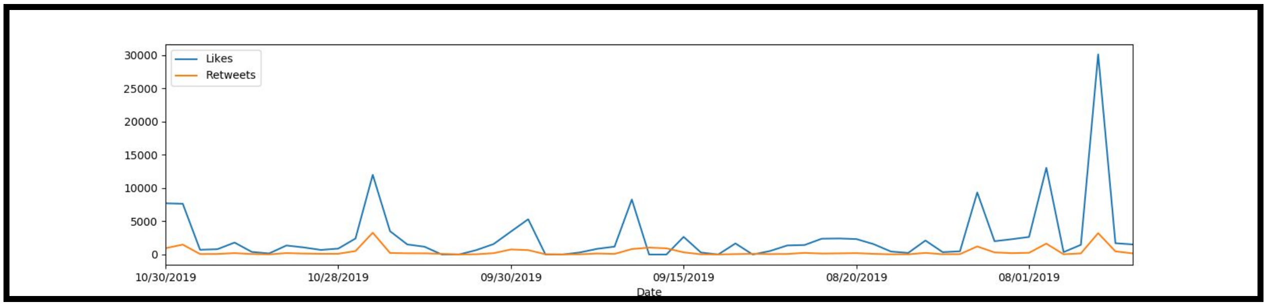
### *3.1 Visualization Insights*

In terms of the visualizations, the time series for the length of tweets clearly shows many dips in length from July to late August. Post-August, his tweets have become noticeably longer as they are consistently reaching the Twitter maximum of 140 characters. Perhaps Piers is becoming more depressed over Arsenal and feels the need to tweet longer rants. On the other hand, the likes vs retweets time series visualization demonstrates that achieving a high amount of likes and retweets is difficult. Piers’ most liked tweet regarding Arsenal came in July at the height of the FIFA Women’s World Cup, where Megan Rapinoe was playing a central role in the U.S.’s winning squad. His most retweeted tweet came last week after Arsenal’s dismal display against minnows Crystal Palace. During this game, Arsenal’s captain, Granit Xhaka, swore at fans, causing Piers Morgan to vent online.

Tweet Length Visualization:



Likes vs Retweets Visualization:



## 4. Reflection

We divided up our work as we went along. Using Professor Li's GitHub guidance on Tweepy as well as the library's documentation, we were able to write the core components of our code. The difficulty arose when we had to integrate the keywords into our analysis of tweets. At times, it was difficult to determine which specific functions each team member was working on. We divided up the work by identifying problems to solve and assigning those to each team member. From there, we would both create separate files to test our code implementation as to not disturb the master code. In hindsight, if this was a more complex project, we could have split up tasks into different code branches and merged them at the end. Another major issue that we managed to solve with Professor Li's help was the speed at which our computers executed the code. By not pickling our data, it took our computers 80-90 seconds to execute the code. Every time we made minor changes to our code, we had to unnecessarily wait to view our output. This made the iterative process frustrating. We wish we had pickled the twitter data from the beginning to save us time and energy. To conclude, this was an enlightening experience for both of us. We learned a tremendous amount regarding the Twitter API, sentiment analysis, data structures, and visualizations. Being able to learn these concepts while working on a project of personal relevance reinforced our learning and will serve as a useful experience in our future coding endeavors and beyond. In today's data and visual driven business world, these skill sets are absolutely crucial when making strategic decisions to solve complex problems. As future business leaders, the two of us could not be more thankful for this exciting opportunity to enhance our professional development.