Sentiment Analysis Surrounding the Creation of the European Super League using Twitter Data

BA 890 – Analytics Practicum Research Paper

GitHub Repository: <https://github.com/AntonioMoralCevallos/BA90-EuropeanSuperLeagueNLP>

Antonio Moral

**Introduction**

This past year has been tumultuous for all, and the football world has not been immune to the effects of the pandemic. Both the 2019-2020 and the 2020-2021 seasons in world football have been like nothing seen before, empty stadiums, cancelled matches, and an overall sense of uncertainty ruled all major competitions. Obviously, this brought enormous financial challenges as a big percentage of earnings for most football clubs in Europe comes from ticket sales to games[[1]](#footnote-1). Amid this, on April 18th 2021, the announcement of a new “European Super League” (ESL), composed of twelve of Europe’s “big” clubs and three invited clubs, caused uproar in the football community. The supposed original purpose of building this new competition was to “provide significantly greater economic growth and support for European football via a long-term commitment to uncapped solidarity payments which will grow in line with league revenues”[[2]](#footnote-2). However, football fans all over the world contested the idea of this elitist league, both online and off, as it was seen as a way to discriminate against smaller teams, and thus take away the essence of “the beautiful game”.

To fully understand the discourse and the sentiment surrounding this announcement, Twitter data was pulled and analyzed using Spacy and other Natural Language Processing (NLP) tools. Through this analysis, the most common topics, organizations, and people mentioned when discussing the European Super League were extracted and interpreted to understand what football fans all around the world thought of this situation.

**Data Scrapping**

To understand the reception of this idea of a new Super League, Twitter was scrapped for tweets that contained the phrase “European Super League” or the hashtag “#EuropeanSuperLeague” and were tweeted out between April 18th and May 20th 2021, as well as in July. These time frames were selected for two main reasons. The first being that capturing the tweets that were emitted right after the proposal would hopefully show the initial reaction of the football world to this news and gives us a broader range of opinions as more people tend to tweet about recent events as they happen. And secondly, gathering data from July 2021 gives us the opportunity to compare sentiments and see how this situation progressed. This is especially interesting considering that because of the backlash received, this Super League was postponed and essentially cancelled, however its effects are still felt in the football community.

The official Twitter API was used to scrap this data, and through it 1000 tweets were extracted from the period between April 18th and May 20th and 300 tweets were extracted for the month of July. It is understandable that there was a significant drop-off in the volume of tweets between the first and the second time period, as again, there are more people talking about a topic more closely to the occurrence of said situation, and especially online, relevance and discourse dies down rapidly[[3]](#footnote-3). Furthermore, although there are likely millions of tweets and reactions that occurred in these two periods, because of limitations imposed by the Twitter API, only a limited amount of data could be explored. However, this should be enough data to uncover the underlying patterns and serve as a good NLP exercise.

A picture containing chart

Description automatically generated

Figure 1 -Searches for the term "European Super League"

**Data Analysis and Result Interpretation**

**Data cleanup**

As with most unstructured data, the tweets pulled had to be cleaned to be able to run a successful analysis. To do this, the NLTK and Spacy modules were used as they provide the necessary tools to turn noisy tweet text data into a usable format. For this project the data was cleaned in two ways, the first being extracting words from tweets individually to then plot the frequency at which these words were used, and the second being cleaning up individual tweets and using this data to extract the entities mentioned as well as applying sentiment analysis. The data cleaning procedure was standard to most other unstructured text data application, consisting of removing punctuation, stop-words and stemming the words in the text. Through this procedure we can standardize and reduce the number of redundant words, making the analysis more efficient. This was especially helpful for the frequency plotting as it decreased the feature space of the data significantly, and it helps condense words with similar meanings together. When dealing with data for entity recognition and sentiment analysis, the words are put into sentences to be able to capture the relationship between words in a sentence, as well as information embedded in the modules used.

**Word frequency comparison**

The first approach to understand the data was to plot the word frequency. Doing so made it possible to see what words were used the most when discussing the European Super League and started showing early signs of the sentiment towards this new idea. Below are the 50 most common words found in the tweets from the announcement period as well as that for the July data extracted.

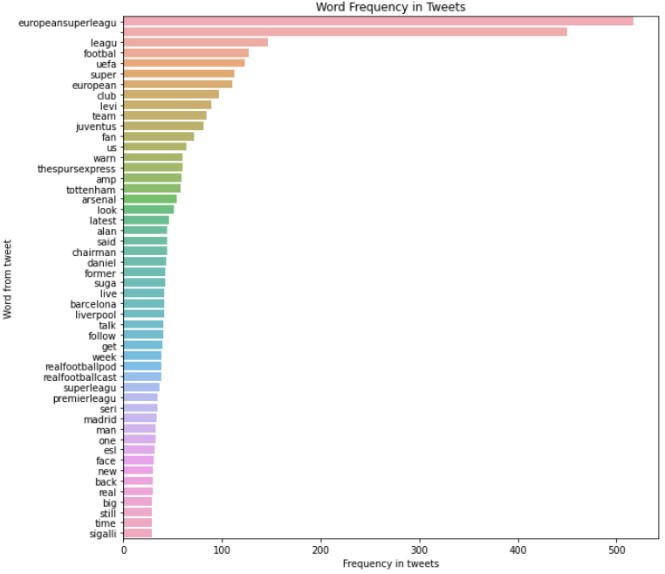
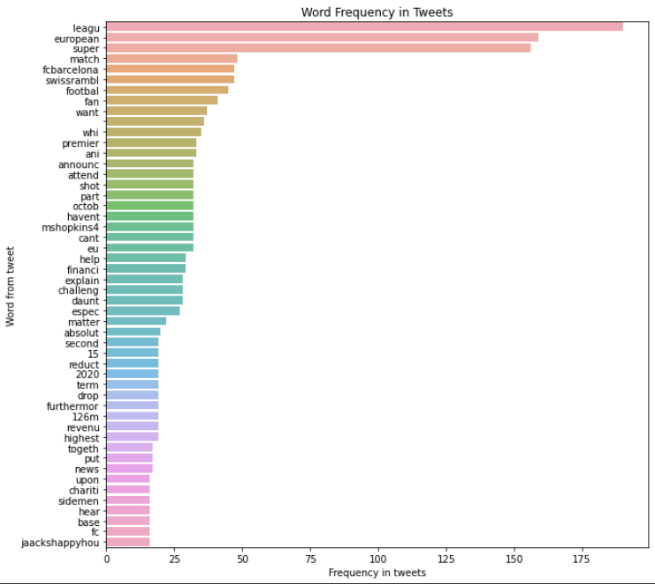


Figure 2 - Word frequency found in tweets from April-May data (left) and July data (right).

From this initial analysis its possible to extract some interesting insights. On the frequency graph for the tweets posted in April and May on the left, we can see that there are more mentions of football club owners, such as Tottenham’s Daniel Levi, and a good amount of the names of the teams that would have been involved in the super league. This makes sense as during this period of time many of the tweets redacted would have been from football news pages reporting on the situation and thus would have reported the teams involved in the to-be scandal. Furthermore, it is interesting to point out that some club names are seen in both graphs, such as for FC Barcelona. This appearance in the second graph could be due to further financial challenges facing the club going into the 2021-2022 season which potentially sparked discussions of whether a concept like the super league was necessary to alleviate said challenges. It is also evident that the data from July tweets is noisier than that for April and May tweets, which is to be expected as the conversation surrounding the ESL diminished with time, and likely the hashtag is added to tweets that do not necessarily discuss the issue. Moreover, it is important to note that as the data scrapped was solely for tweets in English, these results will be biased towards English clubs likely, as people tweeting in other languages were not captured in this study.

**Name Entity Recognition (NER)**

The next step in the analysis was to perform NER on the data to further segment the data and extract more insights. Name entity recognition, as the name implies, consists of the classification of entities in the text through word embeddings. Spacy possesses a powerful method to classify entities in text, and thus that was used for this project. After running both datasets through Spacy’s pipeline, it outputs a labeled set of data with the entity and the category label. This analysis confirmed some of the insights explained above. Mainly, it was possible to observe that the information, and thus entities, found in the tweets from the April-May period were more relevant to the situation and were more interpretable. One reason for this is the noise in the July data, as tweets contain references to entities that are not related to the ESL. Although helpful to investigate, the results from the NER analysis were underwhelming, and very similar to those obtained from the word frequencies.

**Polarity and Subjectivity**

Finally, a polarity and subjectivity analysis were performed on the extracted tweets to understand the sentiment towards this announcement in an analytical way. Spacy possess special methods that automatically calculate the polarity and subjectivity of sentences, and in this case, tweets. This analysis led to some interesting findings when comparing both time periods at hand. After running both datasets through Spacy’s model and aggregating the data, we discovered that the tweets posted in July have a lower polarity score (0.0299) than that for April and May data (0.1052). At a glance, this seems contradictory to the reaction observed in social media in the period after the announcement. Therefore, the combination of polarity and subjectivity was explored to see the whole picture and better explain the average polarity scores for both data sets.

Chart, bar chart

Description automatically generated

Figure 3 - Sentiment Analysis of April-May Data

Chart, histogram

Description automatically generated

Figure 4 - Sentiment Analysis of July Data

Initially, it is evident that both datasets mostly have tweets with a neutral polarity. This is expected as there are multiple tweets posted by sport news sources, and thus are more likely to be unbiased. Moreover, one can see how objective tweets tend to be more positive towards the ESL, which would support the finding that tweets closer to the announcement had a higher average polarity as there were many more news outlets reporting on this situation during that time. However, it is surprising to see that, especially in the April-May data, how there is a noticeable majority of positive subjective tweets according to this analysis, but further investigation into the tweets that were being classified as positive or negative shone a light on one of the biggest difficulties of working with text data. Take for example the following tweet from the April-May dataset: “@leone\_rotich: Arsenal entering the super league 😂😂 #theboyzclub #SuperLeagueOut #SuperLeague #europeansuperleague”, which was classified as a positive tweet by the algorithm. Because of the context surrounding the situation, and more importantly the second hashtag used, it would be seen as critical of the creation of the super league, however, because of the use of the word “super” the model considers this a slightly positive text. With this in mind, and after observing several other similar examples, it is possible to state that this model is not performing well when classifying tweets as positive, especially those with sarcastic undertones or with negative sentiments “hidden” within the hashtags. It is important to note that the model does perform well when classifying negative polarity tweets.

After seeing the low performance classifying positive polarity tweets, the word frequency within negative tweets was analyzed to see how this differed from the overall word frequencies, and this helped us understand further about how the model classified negative tweets, and more importantly what the tweets critical of this new project discussed. This frequency graph for April and May data can be seen below.

Chart, bar chart

Description automatically generated

Figure 5 - Word Frequency for Negative Tweets in April-May Data

Not surprisingly, the most frequent words for this subset of tweets are noticeably more negative than for the overall dataset. From the initial response observed, both online and offline, it seems to accurately represent the sentiments towards this idea of the Super League. It must be pointed out that this is only data from a very small portion of tweets.

**Conclusion**

This project provided insightful findings about both the data at hand and the tools used in the analysis. Through this analysis it was possible to further understand the situation surrounding the announcement of the European Super League and the reaction football fans had in response to it. Overall, it was evident that there was a negative reception to this idea, which is a finding that is supported by experiential research, and it was interesting to observe how a majority of football fans took to criticizing the individual clubs involved, which in the end is a huge reason why this did not transpire further. Moreover, it was a great opportunity to observe how quickly online discourse appears and disappears, considering the magnitude of the creation of such a league.

Although the project was successful in confirming the initial hypotheses, there is a variety of ways to improve both the approach taken and the tools used to develop a more effective analysis. Firstly, one of the biggest hurdles for this project was the limited amount of data that could be extracted from Twitter. After some discussions with Professor Tibert, I discovered that the most efficient way to mine large amounts of data from Twitter is to scrape continually as tweets are being posted, as Twitter is protective of historical tweets. This will be put in practice in future projects and having more data to work with would have helped gain more accurate insights for this project. Considering also the limitations found with Spacy’s base algorithms and models, having more tweet data would have enabled building and training a custom NLP model that better classified the polarity of tweets. In the future, a way to further this analysis can be to, with a personalized NLP pipeline, obtain the word embeddings for tweets, and perform clustering analysis as a separate way to understand underlying patterns in the data. This would also help categorize tweets by polarity using more complex algorithms. It will be interesting to perform similar analysis on other sporting event announcements to see how they differ, or even revisit this project when a new version of the ESL is inevitably announced again in the future.

1. https://www.statista.com/statistics/722949/gate-receipt-revenues-of-football-clubs-europe/ [↑](#footnote-ref-1)
2. https://thesuperleague.com/press.html [↑](#footnote-ref-2)
3. https://trends.google.com/trends/explore?date=2021-04-11%202021-08-15&gprop=news&q=european%20super%20league [↑](#footnote-ref-3)