IST 652 – Scripting for Data Analysis

Student: Sintia Stabel

**Super Bowl 2023 Tweeter Word Cloud and Sentiment Analysis**

The Super Bowl is the championship game of the National Football League (NFL), which is the top professional football league in the United States. The Super Bowl has become an important cultural event in the United States, with 70 to 105 thousand people attending the event, and an average of 100 million viewers watching the game’s broadcast. The game is also known for its high-profile commercials which are broadcasted for the first time during the game and have become a major feature of the event. In the Super Bowl 2023, the teams Philadelphia Eagles and The Kansas City Chiefs faced off during what became the second most-watched program in American television history.

Given its immense popularity, the Super Bowl 2023 is a great subject for social media analysis using tools such as word cloud and sentiment analysis. Tweeter is a social media platform that allows people to share their views on various topics in the form of short messages. The messages are limited to 280 characters or less and the users can share “tweets” publicly or with specific groups of followers. According to the Sports Business Journal, during the game, 603,000 tweets mentioning the Super Bowl were posted. Following is an exploration of 2 data samples containing 2000 tweets each, using a graphical representation called word cloud, and a machine learning model called sentiment analysis to determine if the overall sentiment, would have been a good predictor for who the Super Bowl winner was. To accomplish this, the answer to the following questions were set as targets for the analysis:

1 – Can the overall word frequency and tweets sentiment correctly predict the Super Bowl 2023’s winner?

2 – Is the choice of query words relevant for the accuracy of the results?

**About the data:**

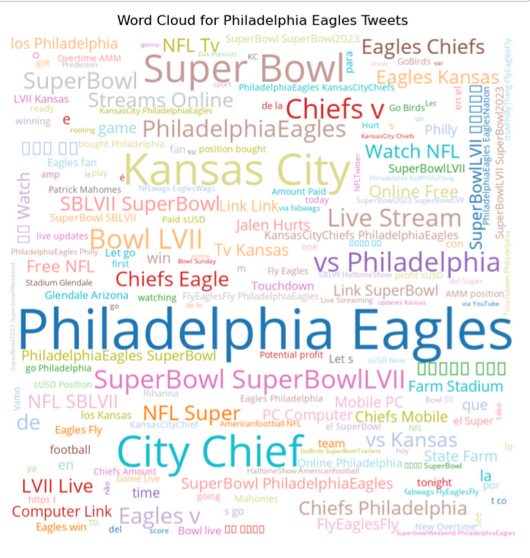
The SNS platform was used to scrape/collect 4000 samples of public tweets on the day of the Super Bowl 2023. The first 2000 tweets were collected using the query “Philadelphia Eagles”, and the second with the words “super bowl 2023”.

The tweets were saved in collections inside MongoDB databases. The collections were then saved as data frames to facilitate transformation and analytics. Preliminary data exploration showed that the “eaglesdb” and the “superdb” data sets, did not require any special transformations for the sentiment analysis purposes, as it would focus specifically on the “content” portion of the data sets. The only transformation performed was the removal of “stop words” prior to running the word cloud analysis.

**1 – Can the overall word frequency and tweets sentiment analysis correctly predict the Super Bowl’s winner?**

To answer this question, a word cloud analysis and a sentiment analysis model were applied to both eaglesdb and superdb data sets.

Word cloud is a graphical representation of text data, where the size of the words corresponds to its frequency or importance within a text. Word clouds are used for visualizing and summarizing large amounts of text data. They provide a quick visualization of the overall theme and sentiment. When applied to the eaglesdb tweets collection, the largest and most noticeable words were: “Philadelphia Eagles”, “City Chief”, “Kansas City”, “Super Bowl”, “vs Philadelphia”, “Eagles Chiefs”, “SuperBowlLVII”, among others. The overall theme appears to slightly favor the Philadelphia Eagles in terms of frequency and importance.



The word cloud for the “super bowl 2023” tweets, produced more even results for both team names, though with Philadelphia having perhaps a little higher frequency. The larger and more frequent words were: “Philadelphia Eagles”, “Kansas City”, “Super Bowl”, “City Chief”, “Live Stream”, “Eagles Kansas”, etc.



Sentiment analysis is a machine learning technique used to identify and extract subjective information from text data, such as opinions, attitudes, and emotions. The goal of the sentiment analysis is to determine the polarity of text data, which can be positive, negative, or neutral. To accomplish this, first the model must be trained to identify positive, negative, and neutral words in order to properly classify them.

For the Super Bowl 2023 tweeter analysis, the TextBlob python library was loaded and used as the polarity baseline. The sentiment analysis results from the “Philadelphia Eagles ”tweets, showed that the tweets were in heavily in favor of the eagles. The chart below confirms this as the frequency for positive polarity is much higher than the negative frequencies. The Eagles were positively portrayed in the collection of tweets collected. The line above zero represents positive polarity while below represents negative polarity.

Chart

Description automatically generated

From the 2000 tweets sample collected using the query words “super bowl 2023”, the sentiment analysis produced a positive polarity regarding the Super Bowl 2023 event. It is important to highlight that the query is asking for tweets that have the words “super bowl 2023”, and not necessarily who will be the winners. So, it does not work in this scenario as a predictor for who the winner would be. Perhaps a sentiment analysis could be used in this context only after isolating the specific tweets that declared who the winners would be. This would require more in-depth Natural Language processing methods, which are beyond the scope of this analysis. Nevertheless, the chart below shows that the positive polarity towards the event far exceed the negative polarity.

Chart

Description automatically generated

**2 – Is the choice of query words relevant for the accuracy of results?**

Yes, the choice of query words played a very important role when collecting samples from the tweeter database. By selecting “Philadelphia Eagles” the results of the sentiment analysis and word cloud were biased towards the “Philadelphia Eagles”. If the query had words like “Kansas City”, we would undoubtedly have seen similar results to what was seen with the “Philadelphia Eagles” query, but with Kansas City as the favorites instead.

When querying using SNS scrape, the SNS will search for tweets that contains all the words in the query in any order. This means that the query will return tweets that mention super, bowl, 2023, Philadelphia, and Eagles anywhere in the tweet, including the content, hashtags, and usernames. Since Kansas City was not part of the query, the results were automatically biased towards Philadelphia Eagles.

The query words “super bowl 2023” produced more neutral results for the word cloud analysis, which proves that without putting a team name specifically on the search, the Eagles were not selected as having most frequent words in the word cloud and were not specifically part of the polarity assessment during the sentiment analysis.

**Conclusion**

The word cloud and sentiment analysis are great tools for quickly assessing the overall sentiment of a large text data sets. As far as predicting who the winner would be, sentiment analysis would have been better suited if the tweets were further filtered to have user inputs that specifically declared their support for a team. Tweeter offers several tears of data that contain user inputs that could have been more appropriate for this analysis but can only be accessed through memberships, and therefore were not used in this study.

The second query produced a word cloud that pointed to the tweets having a lot of promotional content instead of user opinions regarding who the game winners would be. Further investigation of the tweeter feeds and how to optimize query language should be the next steps in improving the results to answers the questions proposed. Furthermore, applying more advanced NLP processing on the content of the tweets could have helped target the tweets with explicit opinions, rather than generalized ones.