Sentiment Analysis Technical Markdown

- Andrew Pierson

The purpose of this project is to gather insight on the popularity of The Kansas City Chiefs during the preseason, and to use this information to predict attendance at future games. This project will provide value to the KC Chiefs organization by benchmarking current popularity and allowing the opportunity to set targets for improving fanbase satisfaction at games. Tweets will be scraped from Twitter to analyze the word counts and sentiments of what Chiefs fans are talking about. A geospatial method will be used to search for tweets containing '#Chiefs" tag within a 25 Kilometer radius of the Longitude and Latitude coordinates of Arrowhead Stadium in Kansas City.

An initial count vectorizer was built using the Twitter text that contained only one parameter which filtered the standard 'english' stopwords list. The count vectorizer tokenized over 1000 terms, many of which were unhelpful for analytical purposes, such as "https", "rt", and "just". A second count vectorizer was built with additional stop words and a min_df parameter that required terms to have a document frequency of at least two percent to be tokenized. This significantly cut down the feature space to just over 100 terms. The resulting list was mostly related to upcoming games and players on the current roster. The third count vectorizer increased the ngram_range parameter, so that only 2- and 3-word terms would be tokenized. This was done with the hope of giving more context to the Tweet text. Players' full names, especially quarterback Patrick Mahomes had many occurences. Some of the most frequently occurring terms included the words, "new currency" and "just dimes", referring to the Chiefs' heavy use of the defensive "dime package".

The porter stemmer and lemmatization were implemented on the Twitter text, however the stemmed and lemmatized words did not provide any further clarity to the analysis. A fourth count vectorizer was created with a higher min_df requirement, filtering words that did not appear in at least five percent of the documents. This was done to filter out the "noise" in the text as was seen in the reduction of the feature space to 33 terms.

After extracting Tweet text and gaining insight into what topics Chiefs fans were talking about, sentiment analysis was performed to gauge the tone of fans' Tweets. Three "dictionaries" were imported to measure whether a Tweet was positive or negative in tone based on its possession of certain words. One dictionary gave different magnitudes of positive or negative scores to words that were positive or negative in nature. Their values were added together based on the words' occurrences in the Tweets, resulting in a positive, negative, or neutral sentiment label if the overall score was greater than, less than, or equal to a chosen threshold, respectively. Another dictionary ascribed positive or negative labels to different words and gave Tweets a positive or negative label based on whether or not there was a majority of words in either direction.

A baseline sentiment analysis was performed in which the threshold for whether a Tweet was labeled as positive or negative was set to zero. One reservation with using this value was that there would be no “buffer zone” for a Tweet to be classified as neutral, which could have resulted in neutral-tone Tweets being labeled as positive or negative. Surprisingly, quite a large number of Tweets ended up being labeled as neutral even without this buffer zone. Upon inspecting the text, many Tweets were objective reports about Chiefs news that contained no sentiment at all. In addition to this, Tweets are very brief by nature which increased the likelihood of a zero-sentiment score. Because of this, the analysis was continued with a cutoff value of zero to determine the sentiment of each Tweet.

Many Tweets were correctly labeled after inspecting the words they contained, such as “winning” and “best” resulting in positive labels. However, some Tweets were mislabeled as negative even though the overall tone was positive. One fan’s Tweet gave praise to Chiefs quarterback Patrick Mahomes, saying his performance made him happy in spite of the team’s poor defensive performance. The tone of this Tweet should have been regarded as positive or neutral, but because the word “bad” was used to describe the defense, and no positive words were used to describe Mahomes’ multiple touchdown passes, the Tweet was labeled as negative.

One challenge in performing the sentiment analysis was that a Tweet’s sentiment could sometimes be gauged by the prevalence of words that were specific to the context of football. In many of the Tweets, “touchdown” or “TD” were used to describe fans’ excitement and praise toward the Chiefs. These words, however, had no positive or negative value attached to them in the sentiment dictionaries. This resulted in some Tweets that had positive tone be mislabeled as negative or neutral. A custom dictionary was created that replaced the words “touchdown” and “TD” with the word “success”. While these words are not necessarily synonymous, inserting “success” into mislabeled Tweets would give more positive weight to the labeling of the Tweets. Doing this did have an impact on the sentiment labels. It changed a negative label to neutral, and three neutral labels to positive.

Based on the observed Tweet text, this method was effective and useful, however this may not always be the case. The word “touchdown” could be used to describe points scored by the opposing team. A fan’s Tweet that expresses frustration about touchdowns scored by the opposing team could be mislabeled as positive if “touchdown” were to be replaced with a positive word.

In this case, the analytical process was manipulated to support a conclusion that was already made about the text. This should only be done with great care and only when it can be clearly seen that the nuances of human language are being misinterpreted by the algorithm. It should never be done to support one’s opinion toward an issue but only to give more clarity to the text. In the instance of replacing “touchdown” with “success,” it is not so much a question of ethics, but rather one of effectiveness and accuracy. As previously discussed, the replacement of words may constructive or destructive to the analysis depending on context of the text. For the Tweets used in this analysis, word

replacement was effective in providing more clarity to the true sentiments of the Tweets. However, this method could easily cause confusion and error to the analysis within a different context. As a data scientist, it is important to be objective in one's approach to the analysis and the data. Thus, manipulation of the analysis should be done very cautiously and sparingly.