CS 5830 - Project 2

Thomas Brower, Zion Steiner

INTRODUCTION

Given football's position in the American cultural pantheon, it's no wonder the Superbowl is celebrated as a pseudo-holiday. On February 2nd, 2020, millions around the U.S. watched the San Francisco 49ers face off against the Kansas City Chiefs. As with all variations of sportsball championship, the stakes are high and the fans are vocal. Many of these fans also share their opinions about the game on Twitter, which serves as a massive repository of material for NLP research.

Our research aims to quantify and contrast the polarity of these opinions between each team's home city. After collecting tweets regarding Superbowl LIV, we conducted sentiment analysis to determine the direction and degree of each city's opinion. Overall, we found that, as expected, the winning team had more favorable tweets in their home city. However, we were unable to determine any prevailing themes or trends present among the tweetset other than the phrases used to search for the tweets in the first place.

DATASET

We used the *Tweepy* Python package to retrieve tweets from Twitter's API. Specifically, we were interested in tweets matching the terms 'chiefs', '49ers', or 'superbowl' made by users located in either San Francisco or Kansas City. The search endpoint's geocode parameter was used to retrieve city-specific search results. Retweets or captioned tweets were ignored. City, message, term that the tweet matched, and time of post were collected for each tweet gathered. Python package *tweet-preprocessor* was used to do tweet cleaning by removing emojis, emoticons, and Twitter-specific substring ex. "RT @".

ANALYSIS TECHNIQUE

After cleaning the data, we completed a sentiment analysis on each of the tweets to identify subjects/themes that were discussed and the strength of the language that was used. We used the package *textblob* to compute a scores measuring the sentiment and subjectivity of each tweet. Sentiment values range from -1.0 to 1.0, with higher values indicating more positive opinions. Subjectivity scores range from 0.0 to 1.0, with lower values indicating more objectiveness. We then subset the tweetset into two groups: subjective (subjective_score >= 0.75) and objective (subjective_score <= 0.25). For each of these groups, a barplot was created showing the mean sentiment score each term received by city.

We also performed Latent Dirichlet Allocation, a method for analyzing topics among a text corpus. This analysis extracted 5 different themes from the tweetset, each comprised of 10 keywords. *PyLDAvis* was used to visualize the relation of these topics.

RESULTS

As was expected, tweets about the Chiefs were generally more positive than tweets about the 49ers in the Kansas City area (*Fig 1)*. What was surprising was that tweets about the 49ers in the Kansas City area were more positive than those from the San Francisco area. This could be a result of Chiefs fans offering their condolences to the 49ers while proclaiming their victory or merely stating the 49ers name at some point during their tweet. This would have skewed the results slightly and is something to consider. It's also worth noting the large confidence intervals in the plot. This indicates that there is large variation between tweets, even for the same term and city.

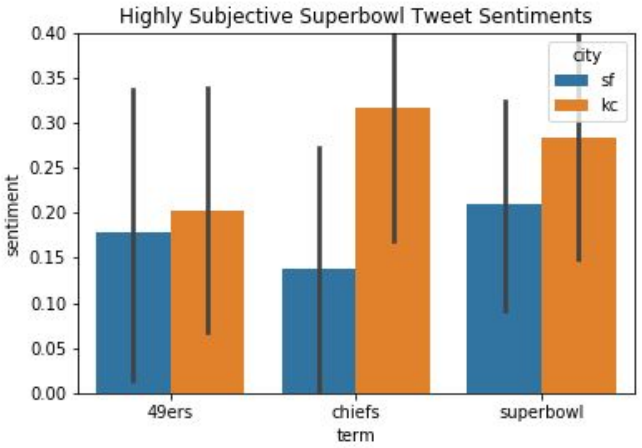

Fig 1. Tweets from Kansas City are more positive overall

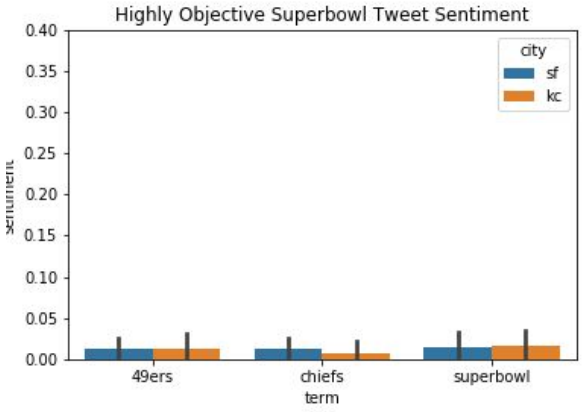

Fig 2. Objective tweets have neutral sentiments and are the same between cities

When looking at the lowest quartile of tweets in terms of subjectivity scores (*Fig 2*), we see that the results are pretty similar across the board regardless of the location being either San Francisco or Kansas City. These tweets are those that are stating that the Superbowl is on, mentioning it in passing, or those that don't have a strong opinion about the Superbowl either way. This is expected as the Superbowl has a lot of people that like it, dislike it, or are passive about its existence. The small confidence intervals on the plot show that objective tweets tended to stay near 0 sentiment, which makes sense given that objective statements are neither positive nor negative.

Finally, when computing the theme analysis we get the following result. We received 5 themes with each theme containing 10 words. The themes were:

Topic #1:
chiefs 49ers superbowl win amp game mahomes city just fans
Topic #2:
bowl super chiefs parade kansas city 49ers tomorrow superbowl win
Topic #3:
chiefs 49ers great team game kansas city superbowl like just
Topic #4:
superbowl 49ers chiefs kc twitter game like team people episode
Topic #5:
49ers coach year reid superbowl andy guys just work head.

Although there are some obvious words that are being shared amongst themes, (49ers, Chiefs, superbowl), we determined that tweets, since they have no minimum requirement, are not long enough to establish a theme amongst one another. *Fig 3* shows that there is some overlap in Themes one and two. Looking at the actual words in Themes 1 and 2 the only words that they have in common are the three words we used to pull our tweets: 49ers, Chiefs, and Superbowl. The strong overlap is due to the frequency those terms are used.

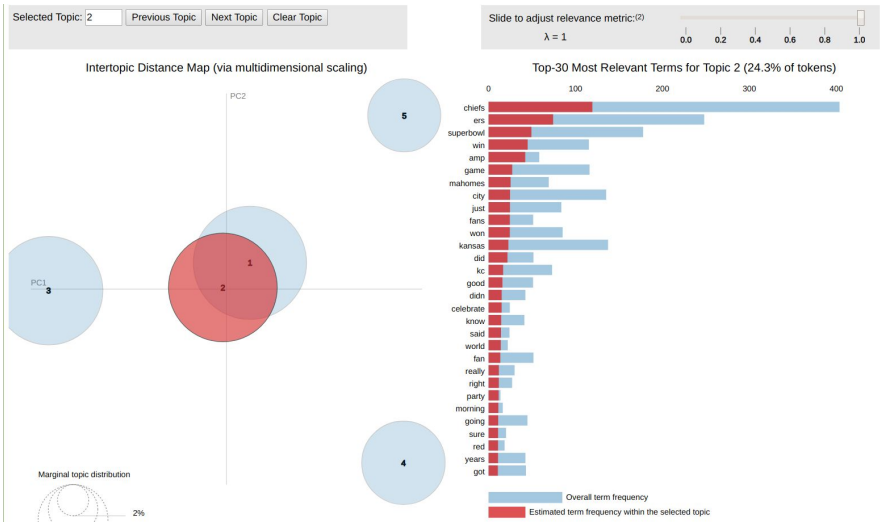

Fig 3. LDA analysis of tweetset produced 5 prevailing topics

While the term sentiment relations present in *Fig 1.* could have been predicted, this research serves as a starting point for answering further questions about the content and evolution of sports-based opinions. For example, conducting this analysis continuously, as opposed to a single timeframe, would allow us to measure changes in the public's opinion on a sports team over time. Augmented with a timeline of team-specific important events, this could give franchise management the ability to measure the effect of each event on team reputation. From a risk-management and marketing standpoint, these are invaluable tools to have for any professional sports management team.