Sentiment Analysis of the NBA and its effect on attendance

Mitchell McCaig

Problem Statement

Is all publicity good publicity? In the realm of combat sports, both negative publicity and positive publicity drive attendance to events. I wondered, does this also apply to team sports? I hope to answer this question through a Sentiment Analysis of tweets prior to regular season games, and whether both negative and positive sentiment drove up attendance.

Background

Understanding the factors that effect attendance can be very powerful for any professional sports team. Putting more fans in the seats will drive the growth of the franchise. Now, with the advent of social media and data science, the sentiment of a fanbase is quantifiable. If a relationship could be found between the sentiment of the fanbase and attendance rates, teams could use this information to forecast future game attendance – fostering further growth. This analysis will be focused on five NBA franchises: the Toronto Raptors, Los Angeles Lakers, New York Knicks, Houston Rockets and Golden State Warriors.

Data

Tweets

Each team's tweet dataset was taken from Twitter using the GetOldTweets3 library. Search queries for the tweets gathered consisted of the team, and the top three players of each team. Tweets were only taken from cities with an NBA team and tweets were taken only during the 2018-2019 regular season (October 2018 - April 2019).

Training Data

Pre-labelled tweets were downloaded from the Qatari Computing Research Institute’s SemEval project (2013-2016) and from the Sentiment140 Project.

Pre-trained word embeddings for the neural network model was taken from Stanford’ GloVe project.

Attendance Data

2018/2019 NBA Season attendance data was scraped from https://www.basketball-reference.com/.

Arena Capacity Data

Arena Capacity Data was scraped from https://en.wikipedia.org/wiki/List_of_National_Basketball_Association_arenas.

Data Cleaning

Attendance Data

Monthly attendance and arena capacity data was scraped using the requests library. The monthly attendance data was then aggregated together into one dataframe to have attendance data for the entire season. Playoff games were removed because games are expected to sell out regardless during the playoffs. Arena capacity was added to the dataframe, and each game's attendance was divided by the arena capacity to find the attendance ratio. An attendance ratio equal to or greater than one denotes a sell-out. Finally, each team's attendance data is filtered and exported to individual csvs.

NBA Tweets

Raw tweets from each team's search queries were then cleaned manually. Because queries such as "warriors" and "rockets" are common words, there are many unrelated tweets within the dataset that need to be removed. Common themes related to the team names were searched and then removed. For example: within the Rockets tweet dataset; "NASA", "Israel", and "SpaceX" tweets were removed – along with many more.

Training Data

The training data taken from the SemEval project was pre-labelled as positive, neutral and negative. Because of class imbalance with negative tweets, additional labelled negative tweets from the Sentiment140 dataset was added to the training set, creating roughly three balances classes. The text data was cleaned by removing: stopwords, hyperlinks, @user information, underscores, apostrophes, hashtags, punctuation, and whitespaces. All letters were converted to lowercase, and each word was lemmatized.

Models

Prior to modelling, the training data is split into training and test sets with an 80/20 split. Both CountVectorizor and TfidfVectorizer are used alongside five other classifiers to compare accuracy and confusion matrix results. In the end, TfidfVectorizer paired with Logistic Regression was chosen to be the predictive model due to high interpretability and all the models have roughly the same accuracies. Once the model was trained, I looked into what tweets were the model not confident in classifying (having a probability < 40%). Those tweets were then removed from the training set, and the model was re-trained. Once re-trained, the penalty and C-value hyperparameters were optimized with an optimal penalty of l1 (Lasso) and C-value of 1. The finalized Logistic Regression model has an accuracy score of 71% - not bad for a three class NLP classifier!

I also wanted to incorporate a recurrent neural network (Figure 1) into this project. Initially, gensim's Word2Vec model was used to see how the word vectors in my corpus were related to each other. The resulting vectors were strange, which led me to use the 100-dimension pre-trained twitter word vectors produced by GloVe. The training data was split again into training and set sets with an 80/20 split, and the training data was padded, label encoded and split into a train and validation set. Finally, the 100-dimension word embedding was converted into a matrix and place into the first layer of the neural network. The neural network architecture consists of a 100-dimension word embedding layer, LSTM layer with relu activation and a Dense layer that classifies the tweet. The neural network achieved an accuracy of 72% - only marginally higher than the simpler Logistic Regression model. Because of this, the Logisitic Regression model was used to perform the sentiment analysis predictions.

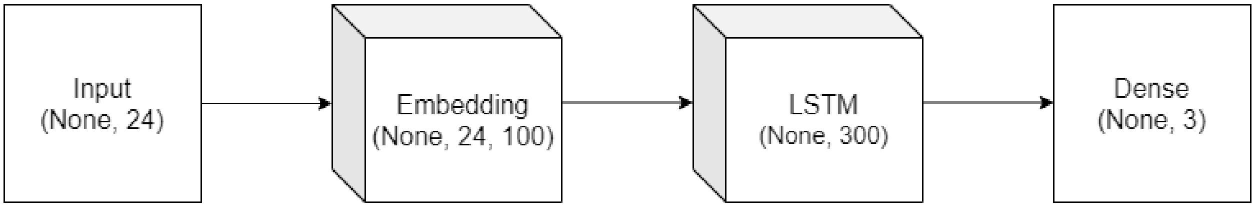

Figure 1 - RNN Architecture

Analysis

Each team's NBA tweets had their sentiments predicted and labelled using the Logistic Regression model. I then proceeded to calculate the ratio of polarized tweets to the total amount of tweets prior to a game. My hypothesis being that both negative and positive sentiment will drive attendance to the game, negative and positive tweets were bundled together as polarized tweets. The polarized sentiment ratio was then plotted along with the attendance ratio to see if there was any linear relationship.

Conclusion

There is not enough evidence to conclude that tweet sentiment prior to games influences attendance for any of these teams. Games for most of these teams sell out during the regular season, regardless of the sentiment on twitter (Figure 2), resulting in no seen linear relationships. Additional graphs can be found in the GitHub Repository.

Future Analysis

Future analysis for this project will be focused instead on season-to-season attendance rather than game-to-game. Professional teams do suffer from lower attendance rates if the team suffers from consecutive losing seasons, and I hope to quantify this using sentiment. I would like to add in more teams as well, especially smaller market teams as they should also be more sensitive to sentiment affects. Finally, incorporating other social media platforms such as Reddit or Instagram comments will add more information to the analysis.

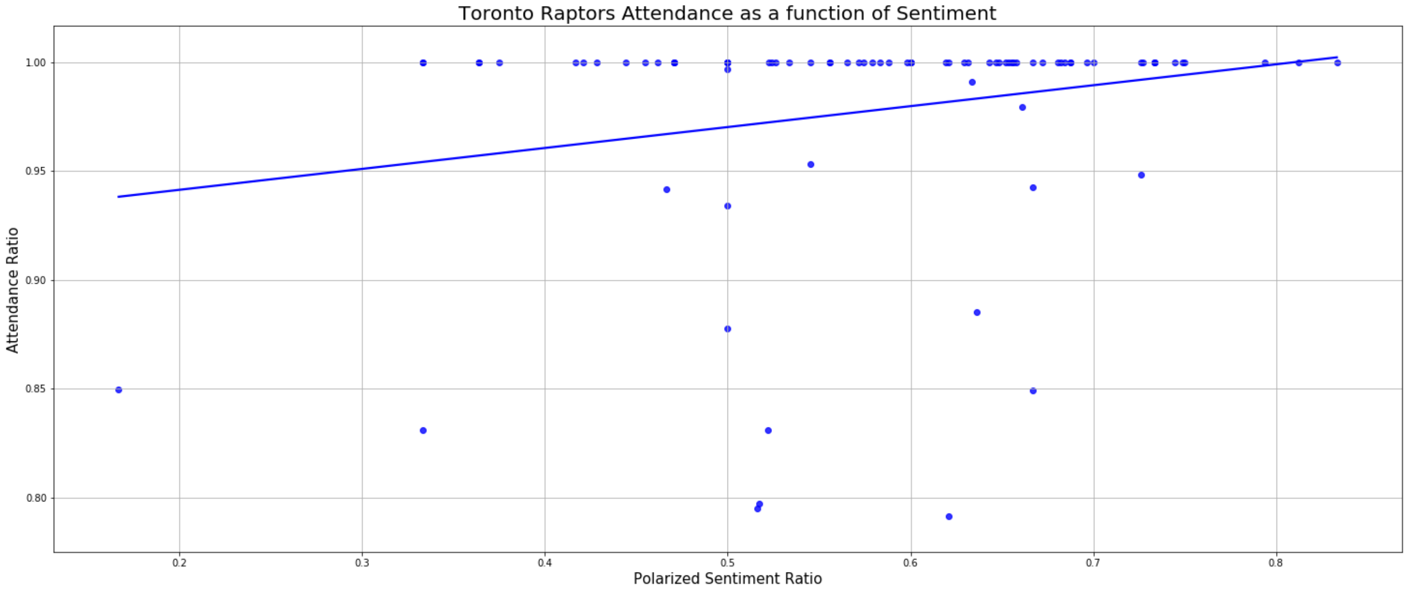

Figure 2 - No linear relationship between sentiment and attendance for the Toronto Raptors 2018/2019 regular season