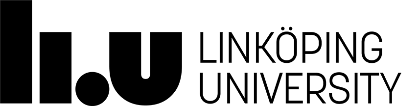
**Sentimental Analysis On Tweets of National Hockey League Teams**

Fahad Hameed

fahha780

Text Mining

732A92

**Abstract**

*In this project, Social media site Twitter is used to analyze National Hockey League (NHL) Teams Popularity, for which teams related tweets are used to do sentiment analysis. Determining which teams’ popularity is increasing or decreasing and public reviews with the period of time on twitter will help sponsors of NHL to decide on which team to invest. Machine leaning classification models Naïve Bayes Classifier and Deep Neural Network developed in the project for classifying tweets can be used for classifying future tweets and determine the distribution of positive and negative tweets. Using team social media sentiments together with teams on field performance can be used to rank teams.*

**Index**

[1. Introduction 4](#_Toc508806923)

[2. Theory: 4](#_Toc508806924)

[Textblob 4](#_Toc508806925)

[Vander-Sentiment Analysis 5](#_Toc508806926)

[Naive Bayes Classifer 5](#_Toc508806926)

[Deep Neural Network 5](#_Toc508806926)

[3. Data of the project 6](#_Toc508806927)

[Teams Names 6](#_Toc508806929)

[Tweets 6](#_Toc508806930)

[Data Preprocessing 6](#_Toc508806932)

[4. Methodology 7](#_Toc508806933)

[Data Retrival 7](#_Toc508806934)

[Data Preprocessing 7](#_Toc508806935)

[Data Analysis 7](#_Toc508806936)

[Sentimental Analysis 8](#_Toc508806936)

[Classification Modeling 8](#_Toc508806936)

[5. Results 9](#_Toc508806937)

[6. Discussion 10](#_Toc508806938)

[7. Conclusion 11](#_Toc508806939)

1. **Introduction**

Today social media is an important part in sports, people use it to know what is currently going on and what the future events are. Along with that it is also used by the teams or players to promote themselves and connect with their viewers. Almost every team, league and sports association has a social media profile on Twitter. From the pros to the minors and from the high school athlete to the retired athlete, social media has been a force in the sports industry landscape.

In any sports sponsors are integral part who support their teams financially. It is crucial for the sponsors to understand well teams performances along with their popularity. This may result in more audience in matches, better sponsors, and more money to fulfill and increase objectives of the teams. In this project I target one social media site Twitter to measure the popularity of the teams in National Hockey League (NHL). NHL seasons occur regularly which makes it necessary to know that which team is more popular in the current season. Also to see what is the reason of team’s popularity ranking and is there any relation with the performance of the teams in the field or there are some other things that are reason of it irrespective of their field performance. In the project I tried to find out what NHL twitter followers are thinking about teams. Getting to know which teams are popular or not will eventually help teams to improve their performance and get a view what they need to do in order to improve their public perspective. For sponsors it is necessary to finance team which is more popular and have positive viewer ship for their better brand promotion. As a sport enthusiast I chose this project so that I can better understand the impact of social media on sports as social media is a powerful tool today on which sports talk happen and the way fans interact with teams, players, personalities and fellow fans.

1. **Theory**

Text Classification in Natural Language Processing (NLP) is a technique where we assign one or more categories or classes to the document to classify it into different types of documents. As there is a lot of content on the web there is a need for classifying and understanding the text. Social media data is increasing vigorously with the number of internet users as said in 2018 there were 3.196 billion users. It is important for social media analysis to understand it in some automated manner. There are many tools for this purpose but I am presenting a few which are used in this project.

**TextBlob**

Textblob is a natural language processing library for textual data which is built on top of NLTK. It helps in common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, Tokenization, Lemmatization, WordNet Integration, n-grams and more.

**VADER-Sentiment-Analysis**

VADER-Sentiment-Analysis is a lexicon and rule based sentiment analysis tool that is specifically designed for social media. Vender Sentiment typically deals with sentences considering typically special characters, negations, use of contractions as negations, conventional use of punctuation, word-shape all capital letters, using utf-8 encoded emoji’s, initialisms and acronyms to understand sentence sentiments, which are very important in sentences to express emotions. It assign polarity to sentences with +1 for positive sentence, zero for neutral and -1 for negative sentences.

**Naive Bayes Classifier**

Naive Bayes is a useful technique to apply in text classification problems which is based on the bag-of-words model. It uses the probabilities of observed outcomes to return a reasonable estimate of an unknown outcome. In bag-of-words model word appearing in a positive words list will cause total scores to update by +1 or if it is appearing in negative word list score is update with -1. Depending on the total score text is classified as positive or negative text. The probability a document belongs to a class ***C*** is given by the class probability ***P(C)*** multiplied by the products of the conditional probabilities of each word for that class.

C:\Users\Fahad Hameed\Desktop\NaiveBayes-1.PNG

***Count(di,C)*** number of occurrences of word ***di*** in class ***C***,

***VC***  total number of words in class ***C***

***n:*** number of words in the document we are currently classifying

**Deep Neural Networks**

Deep Neural networks are complex structures organized into input, hidden and output layers. NN learn by comparing their classification of the record with the known actual classification of the dataset. The errors from the initial classification of the first record is fed back into the network, and used to modify the networks algorithm for further iterations. The deep neural network structure I used in this project is having four layers.

**Embedding Layer**

**(10000, 100)**

**Dense Layer**

**LSTM**

**Dropout Layer**

**OUTPUT**

**INPUT**

The **Embedding Layer** expands each token to a more massive vector, which allow the network to represent a word in a meaningful way, Layer is getting an input which initialize random weights and learn embedding for all the words in the training dataset, 10000 size of the library and 100 which is dimension of the embedding. **Dropout layer** is a regularization layer where randomly selected layers are ignored. Long Short-Term Memory (**LSTM**) networks are recurrent neural network having capacity of learning order dependence in sequence prediction problems which are widely used in language processing models. The last layer in are DNN is **Dense layer** which is using *softmax* as an activation function with 2 as shape of the output.

1. **Project Data**

Today social media is an important part of sports almost all sport teams, leagues and players have their social media presence especially on twitter. *According to*[*Navigate Research*](http://www.navigateresearch.com/)*, sports fans are 67 percent more likely to use twitter to enhance their viewing experience compared to non-sports fans*. Twitter is a platform where fans and players communicate before and after the matches to share their thinking. So for the analysis of NHL I considered twitter as the best platform and most relevant. This project consists of two data extraction parts which are:

1. **Team names:**

The first part of data extraction is getting the National Hockey League (NHL) team names and rankings from the ESPN website. For this *beautifulSoup* library was used and 31 team names were extracted and stored in a csv file.

1. **Tweets:**

Second part consists of retrieving Tweets related to NHL teams with hashtag of team names including Author name and date of tweets. *Tweepy* library was used to get the latest tweets from twitter. Retrieved tweets were stored in separate csv file. I collected tweets for each team separately to get specific results for each team and also to get more number of tweets. For each team 500 tweets were collected and for 31 teams 15500 tweets. Along with it separately tweets for different hashtags like "NHL", "NHL19", "NHL2019", and “National Hockey League" about NHL collected to know the general trend of NHL team’s popularity.

**Data Preprocessing**

Tweets collected were first preprocessed before analyzing the tweets data statistics and applying the machine learning models. I used the following data cleaning and processing steps to the tweets.

**Tweet Cleaning:**

1. Removing unnecessary URLs.
2. Removing hashtag (#) and user mention (@) symbols.
3. Removing punctuation marks, only specific punctuation marks were removed as they are very important in expressing the emotions of the users.
4. Converting letters that appear in repetition to the specific word.
5. Removing invalid words, converting words to lower case and handling retweets.

These two functions were used for the data cleaning:

* *strip\_links()* removing websites links from the tweets because they are not useful for sentimental analysis.
* *strip\_all\_entities()* Removing punctuation marks from the text like hashtags and at the rate symbol

**Tokenize**: Tokenization of the tweets into words, keywords, phrases, symbols called tokens using nltk.tokenize library.

**Stemming**: Words in tweets were stemmed i.e. reducing infected words or derived words to their base form.

**Stop words Removal**: Words in tweets that do not contain any meaning information were removed using python library *stopwords* for English.

*tokenization\_and\_stem()*function was used for the purpose of tokenization, stemming and stop words removal.

**Joining:** At the end again words were joined to make meaningful tweets by using space. *token\_join()* function was used to convert tokens to sentences.

1. **Methodology**

The project consists of five parts; Data retrieval, Data Cleaning, Data analysis, Sentimental Analysis and the last part machine learning classification modeling.

1. **Data retrieval:**

First step in the project is the collection of data which consists of NHL teams and the relevant tweets. For data extraction the “Data Extraction.ipynb” file is used and all the data is stored in separate csv files for all teams which is explained above in Project Data part.

Next two steps are performed using “Data Explore Sentimental Analysis.ipynb” file is used.

1. **Data Preprocessing:**

After processing tweets using the above mentioned process in data preprocessing section next step of data analysis was performed.

1. **Data Analysis:**

In thispart statistics about the data are analyzed using *item\_count()* function which gives information about number of words in tweets, Mentioned number of user, Hashtags, URLs, special characters and emoji. Also the most used words in the tweets. Some of the statistics found by performing it on 15500 tweets are presented below:

|  |  |
| --- | --- |
| Number of users Mentioned in the tweets | 15119 |
| Same user Mention Maximum number of times | 49 |
| Number of URLs | 8954 |
| Emojis | 40 (Positive: 12 Negative: 28) |
| Number of Words | 273574 |
| Unique Words | 10904 |
| Bigrams | 258077 |
| Unique Bigrams | 60753 |

Word Frequency Distribution was developed to see which words are mostly used by the people. For this stopwords and some other words like names of teams etc were removed as they will appear frequently in the dataset. Some most common words found were “game”, “fan”, “help”, “win”, “tonight”, “goal”, “play” and “need”.

1. **Sentimental Analysis:**

Sentimental analysis of the tweets is done using functions sentiment\_analysis() and assign\_sentiments(). These two functions uses Vander-Sentiment Analysis for assigning sentiments to the already cleaned tweets. Tweets with sentiment polarity greater than 0.005 are assigned positive sentiment and less than -0.005 are assigned negative sentiments and others with neutral sentiment. It stores tweets with the sentiments in a new csv file.

These functions were applied on Original, Cleaned, Stemmed and tokenized tweets to get a view of what difference we get by using these processed tweets. For cleaned tweets following sentiments were assigned:

* **Positive Tweets:** 7238
* **Negative Tweets:** 3125
* **Neutral Tweets:** 5137

All the four got almost same result with a difference of 10 to 20 tweets sentiments. I used cleaned Tweets for further process of classification modeling.

1. **Classification Modeling:** Two classification models are used in the project for assigning sentiments to newly collected tweets which are Naïve Bayes Classifier and Deep Neural Network. “Model.ipynb” file is used for the classification. It uses tweets with assigned sentiments in the previous step for training, testing and validation. As I was only interested in the positive and negative tweets so neutral tweets were removed. When feeding data to the classifiers equal amount of positive and negative tweets were used so that the classifiers learns both of them equally, despite the number of positive tweets were much greater than the negative tweets. Supplying unbalanced tweets results in wrong prediction or giving positive prediction more value.
2. **Naïve Bayes classifier** is implemented in class *Naïve\_Classifier* which have three functions fit, predict and score. The *Naïve\_Classifier* class takes training and test dataset I used 25% of the data in Naïve Bayes classifier as test data and rest as training data. Testing the score of the model with score function of class resulted in 91% accuracy.

**Deep Neural Network (DNN**) with four layers is implemented performing the following steps:

Firstly text is converted into sequences of token using Tokenizer, fit\_on\_texts() and texts\_to\_sequence() functions.

I used Stanford *glove.twitter* which is pre trained model on Twitter posts and can be fed to the neural network. Using Glove improved the accuracy to 3%.

In the Neural Network model I used 30% of data as test data and rest as training data and validation which resulted in 93% accuracy.

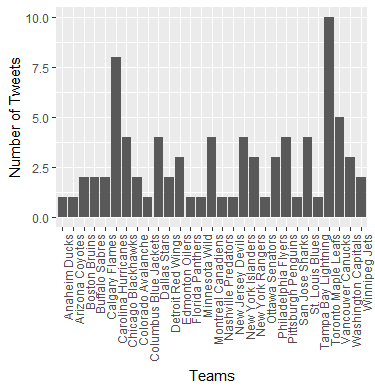
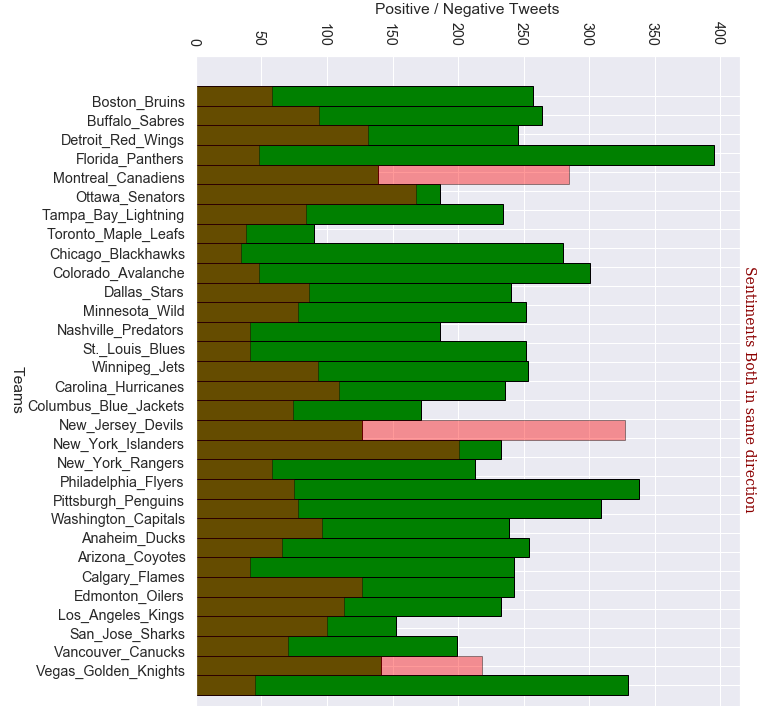
1. **Results**

The sentiments about the teams and the ranking called Total Points (PTS) for every team from ESPN website shows quite good relation between the two. Eight teams with top and worst rankings are show in the tables below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Team Names** | **ESPN Ranking (PTS)** | **Positive Tweets** | **Negative Tweets** |
| 1 | Tampa Bay Lightning | 118 | 234 | 84 |
| 2 | Boston Bruins | 97 | 257 | 58 |
| 3 | Calgary Flames | 97 | 243 | 127 |
| 4 | San Jose Sharks | 94 | 199 | 70 |
| 28 | New Jersey Devils | 63 | 127 | 327 |
| 29 | Detroit Red Wings | 62 | 246 | 131 |
| 30 | Los Angeles Kings | 58 | 153 | 100 |
| 31 | Ottawa Senators | 56 | 186 | 168 |

*Table 1*

From comparing the bar chart below and the ranking it was evident that for the teams whose performance was worse received more negative sentiments as compared to teams who performed well received positive sentiments.



*Figure.1 Teams Positive Tweets in Green and Negative Tweets in Red*

*Figure.2 Teams Trends*

Other objective of the project to create classifier which can be used for future tweets for sentimental analysis instead of using sentimental analysis tool every time and improve the prediction of sentiments; Naïve Bayes classifier got an accuracy of 92% and deep neural network an accuracy of 93%.

1. **Discussion**

The results from Table 1 show relation between the ESPN ranking and the public tweets sentiments, positive/negative tweets proportion when compared with the ranking it is evident that top ranked teams have quite large number of positive tweets as compared to negative tweets and for worst ranked teams they have less positive tweets as compared to negative tweets. Figure 1 shows positive and negative tweets about each team, fans mostly like to talk positive about the teams and support them. This is something that is understandable that if the teams perform well there is are positive comments about the teams otherwise negative. The model helps to understand the activity of users. Three teams have much more negative tweets Montreal Canadiens, New Jersey Devils and Vancouver Canucks. The latter two teams have very low ranking in ESPN Third last and Sixth last respectively, so it is understandable that negative reviews are due to bad performance whereas Montreal Canadiens have average ranking, It may be due to some other reason other than on field performance as social media perspective about teams can be due to player behavior or other incidents. But the general behavior of the sentiments which were received for the teams and the ESPN ranking with which I compared the results are quite relevant.

Figure 2 in the results section shows trends about which team most people are talking. This data was collected using different hashtags about NHL and it shows Toronto Maple Leafs, Carolina Hurricanes and Vancouver Canucks as most popular teams. These teams are not even in ESPN ranking or do not have high positive or negative sentiments which is quite different from the analysis what was made when looking at teams individually. Which maybe be due to some other reason other than team performance on the field.

The two classifier are used in the project for classifying tweets into positive or negative so that every time there is no need to use vender sentiment analysis tool for assigning sentiments of our tweets instead we can use the trained models to assign sentiments to upcoming tweets on regular basis.

The problems that were faced during the project are the following ones:

Due to Twitter API limitation limited amount of data can be collected at once. API also limits tweets collection which are older than 20 days. Getting more data could have increased the accuracy of models.

In the project I only considered twitter sentiments and compared it to ranking which is completely based on teams on field performance. The classifier can show the general public perception about the teams but it cannot be solely compared to on field ranking.

1. **Conclusion**

In the project from results and their analysis it can be seen that the social analysis of National Hockey League Teams on twitter can be really helpful to decide about some key thing like which team to sponsor, offer advertisement and also need to improve public interaction for the teams. Getting data on regularly basis and analyzing trends can be very helpful to understand the reason for positive and negative popularity of teams when compared to the ranking of the teams.

During the project I learned many technical things along with how to work on a project individually in a systematic manner starting from scratch to graphical results. I gained valuable experience which will help me to improve my Natural Language Processing knowledge and add more to my profile. Going through the process I learned how to scrap data from websites and also get data using API, manipulating twitter data to make it more reasonable and do data cleaning on it. Doing sentiment analysis using different types of preprocessed data. Understanding statistics of data using different visualizations. One of the important thing which I learned is how to use machine learning models to classify tweets and improve the model to make it better. I used TensorFlow and Keras for neural network which gave me new knowledge about these popular python libraries. Using different layers in neural network model and playing with them to improve model, increased my machine learning concepts about LSTM and regularization.

**References**

* https://www.forbes.com/sites/anthonydimoro/2015/07/02/the-growing-impact-of-social-media-/-todays-sports-culture/#7ad52a8931a2
* <https://www.tensorflow.org/tutorials/keras/basic_text_classification>
* https://devklaus.wordpress.com/2017/09/28/sentiment-analysis-to-find-problems/
* https://towardsdatascience.com/do-tweets-from-nba-leading-players-have-correlations-with-their-performance-7358c79aa216