3-22-2021

The Impact of Twitter on the National Hockey League and its Players

Benjamin Strauss

Bryant University

HONORS THESIS

The Impact of Twitter on the National Hockey League and its Players

BY Benjamin Strauss

ADVISOR • Kevin Mentzer
EDITORIAL REVIEWER • Suhong Li

Submitted in partial fulfillment of the requirements for graduation
with honors in the Bryant University Honors Program
April 2021

TABLE OF CONTENTS

ABSTRACT

This study offers a new perspective on collecting and analyzing Twitter data surrounding the National Hockey League (NHL) to identify any trends or relationships between the data and overall performance during the 2021 abbreviated season. This paper provides and in-depth analysis by studying a sample of sixty of the top NHL players, specifically those who are typically top performers in the league, spanning over all thirty-one teams and all positions, this study was able to identify a deeper and broader perspective of what implications can be drawn from analyzing data from Twitter to both predict and reflect both individual player and team performance. In using a set of identified statistics to study performance, as well as analysis techniques, primarily focusing on sentiment and volume to study the Twitter data, this paper defines key relationships which can be used to draw future implications on performance. This paper combines multiple tests and models incorporating platforms and programming languages including Python, Visual Basic for Applications, and Structured Query Language (SQL) to identify trends in these leading players and teams in order to identify if there are any predictive or reflective features which can be identified when comparing Twitter data to performance. My analysis is highly relevant to the NHL and can be replicated across many other teams, leagues, and platforms as they can benchmark these results against one another, in hopes that players, coaches, analysts, and viewers can all benefit from these findings. In this study, multiple tests are conducted to find the characteristics of the most and least successful players and teams within the league. The results of this study show clear indications of my initial predictions of this study: a positive relationship exists between the Twitter data, specifically volume and sentiment, and performance among players and teams.

INTRODUCTION

The purpose of this study is to find if there is a relationship between social media usage, primarily Twitter, and performance among the National Hockey League (NHL). I will be studying the frequency of usage, the overall temperament during usage, as well as social media use from the public targeting professional hockey players in order to find if social media usage is predictive or reflective of performance, or potentially both. I will be comparing the usage of social media and performance during the abbreviated 2021 NHL season, starting two months before the start date, and going through the first eight weeks of the season, against the usage and performance of a sample of sixty of the highest performing and generally most popular NHL players, ranging across all positions and teams in the league.

This topic is relevant and important to other scholars or researchers in the field because my study will provide information on the relationship between these aspects in attempt to help teams, coaches, and players predict performance and help alter team and individual strategies as a result. Hopefully, from my research and analysis, team and league management will be able to predict how certain players will perform and how to alter game strategy as a result. I believe that my results will relate strongly to all levels of hockey, so the NHL, and other leagues, could use this dataset and study to help with their performance analyses. Additionally, management could begin set restrictions in contracts involving usage of social media if the correlation is strong. Ultimately, I am hoping to discover a new predictive factor that can be utilized by anyone hoping to track and predict performance.

I am attempting to find if there is a relationship between social media usage and performance and what that relationship is. I am trying to find if the correlation— assuming one exists— will be useful in predicting outcomes, performance, and strategy. Additionally, I am trying to see if the overall temperament of the social media post has any correlation to performance. Prior studies have noted that there is a negative correlation between social media usage and overall performance in professional sports, so I believe that I will find the same relationship in NHL, but I would like to get a more in depth understanding, as well as attempt to see if social media usage can predict performance, as opposed to just reflect.

This study will be a research thesis consisting of observational data that I will collect from Twitter surrounding the NHL season, using Tweets from November 2019 through March 2021. The Tweets will be collected based on key words, accounts, and hashtags relating to the sample pool of players and the thirty-one NHL teams. With the data from the league, specifically the sixty players that I chose to analyze, as well as each of the teams, I will run various summary statistics and tests to clean the data and then visualize it in a clear and concise way. I will use programs including Excel, Tableau, and Python to complete my project and run my analyses.

In the past there have been a decent number of studies completed about the impact of social media and performance. However, these projects typically do not revolve around team sports, rather just individual sports such as golf or bowling, and none revolve around the National Hockey League, which could be attributed to the fact that NHL players do not have a large presence on social media compared to sports like football and basketball; however, there is plenty of attention and activity towards the league from external sources including the fans, analysts, and media. Furthermore, these studies do not focus solely on Twitter usage, just social media in general; therefore, my project is unique in both of these aspects. Typically, these studies track heavy social media users in professional sports, to see if their usage reflects performance. For the most part these studies have shown that higher social media usage typically reflects lower performance. Although, I would like to find out if this test can be used to predict performance, as opposed to just reflect it. The goal of this study would be to give players and coaches a resource and better understanding on how social media can impact a player or team. If the relationship is strong enough, coaches and managers may use these studies to consider limiting or encouraging social media usage among their team.

I chose to use Twitter because its sentiment data is easily accessible and easy to analyze. There is no need to conduct a study involving questionnaires or survey and all the data are free and perfectly legal to collect. Furthermore, there are a massive amount of data accessible and relevant to this study that I was able to incorporate to my advantage. There are still limitations when using Twitter, but using this dataset captures a lot more voices and perspectives.

Using Python, I was able to look at the Tweets over this period of time and find the sentiment score of each of these Tweets. The sentiment score is a numerical value ranging from negative one to positive one which analyzes the words in a Tweet and scores how positive or negative it is. For instance, a Tweet with a score of 0.8 is positive and probably contains positive words such as “love,” “nice,” “great,” and so on, whereas a Tweet with a score of -0.8 is negative and would most likely contain words similar to “hate,” “bad,” “awful,” and so on. From here, I was able to look at the sentiment over the course of the season among each of the players and the teams to provide insight on how these players and teams are being talked about on Twitter, and then look into why this is the case, and whether it can then be attributed to being predictive or reflective of performance. I broke up these sentiment scores by week to get a more in-depth analysis and to be able to pinpoint certain events during the season. Therefore, for example, say a player in my study scores a high sentiment average for a given week, could this potentially be the result of a strong or weak performance the prior week, or in turn, could this predict a strong or weak performance in the next week.

Looking at the sentiment over the course of the season provides good insight into not only the players and teams, but the fanbases and sports analysts too. There is the possibility that the way people talk about a certain player and how frequently can impact or predict how they are going to play in a following game, as opposed to just reflecting or representing a previous performance. This study also provides strong insight into sports leagues in general, as this study can be replicated among various sports and leagues with a social media presence.

This study was originally supposed to focus on the entirety of an NHL season, including preseason, eight-two regular season games, and playoffs; however, due to the COVID-19 pandemic, this season was cut short and delayed a few months. As a result, this study looks at Twitter data starting two months prior to the season start, followed by eight weeks of data from the actual season itself, spanning about half of the now fifty-six game abbreviated season. Typically, NHL seasons start at the beginning of October, but this season began in mid-January and was cut short by twenty-six games. Since I was not able to collect data on an entire and normal season, there are certain limitations to my study. I originally wanted to use my analysis to predict various team, player, and coach awards which are announced following the conclusion of the season. Although I can make predictions to these, I will not be able to know for sure until the 2021 NHL Awards. Additionally, fans are hardly allowed to be in attendance for the games during the season, if at all, so this would be interesting to see if there is any difference in my analysis among a season with no fans in attendance versus a season with full attendance, especially regarding more popular players or teams with a larger fanbase. Finally, the teams were split into four different conferences for this season based on location to reduce travel distance and have the teams play in more a series format, similar to the MLB. Typically, each team would play every team at least once throughout the season, but this year each team plays each of the seven other teams in their respective conference eight times, totaling fifty-six games. Regardless, the reduced season still provided me with the appropriate insight I needed to draw conclusions about how the Twitter world plays a role in professional hockey, but further data could be collected to determine if my analysis holds true for a full eighty-two game season. The significance levels were relatively low as well for my correlation matrices throughout my study, so this is not necessarily an issue, but a limitation, nonetheless. Finally, I wanted to work with multiple different platforms and languages including Tableau, Excel, Python, and SQL, to demonstrate my understanding on each of these concepts, and although I was successful in this, using multiple tools and transferring data between these can lead to human error and potentially impact the study. Thankfully, I had a lot of eyes overseeing my work, but in the future, with a more advanced skillset, I think solely using Python would be beneficial as everything can be done using this. Therefore, a replication with minor tweaks and extensions should ultimately be conducted to confirm my results.

Ultimately, professional athletes are consumed by social media, which can in turn, impact performance. The purpose of this study is to find if there is a relationship between Twitter and performance among the National Hockey League, in order to find if social media usage is predictive or reflective of performance. Twitter sentiment will obviously reflect performance as people constantly respond to games or events on social media; however, I expect that I will be able to find a predictive relationship between performance and sentiment as well.

LITERATURE REVIEW

The main body of literature that I focus on to assist me and my study is titled *Analytic Methods in Sports* by Thomas A. Severini, which uses mathematics and statistics to understand data from baseball, football, basketball, and other sports. There was great insight on various analytics that can be used when studying sports of all levels. There is far more than what meets the eye when it comes to sports analytics, and those that are able to think out of the box and dive deeper into the world of analytics can find more team and individual success. Some of the most successful teams and coaches use analytics and models to their advantage when preparing and developing strategy. This book helped me tremendously in terms of looking at the mathematics and statistics involved in my analysis and how to actually go about analyzing all this data efficiently.

A key takeaway from Severini, was his analysis on average time on ice (TOI) and points scored. Severini notes throughout his study that time on ice must be accounted for as it is not fair to compare players who hardly play to star players. Therefore, adjusting statistics based on regression models, or in my case, sticking to players which consistently play is necessary. I was able to include players who are injured as well to benchmark my results, but in Severini's time on ice versus points scored model, only players playing twenty or more minutes per game in at least sixty games of the season were considered, totaling a sample of three hundred players (Appendix I). Essentially, this study found a clear positive correlation between average time on ice and points scored, and although this study was conducted five or so years ago and a portion of the sample have since retired from professional hockey, the same results would still be found in a more modern pool. A regression analysis, using a linear function, was performed, finding the resulting regression equation: $\hat{y} = -0.505 + 0.0639T$ with an r-squared value of 71.9%, (Severini, 2015, p. 157). However, the resulting plot with the regression line, as shown in Appendix II, "suggests that there is a nonlinear component to the relationship; for instance, the points tend to lie about the regression line for small and large values of [time on ice]," (Severini, 2015, p. 158). Therefore, Severini used a quadratic regression model which provided a more accurate representation of the data at hand, which resulted in the regression equation: $\hat{y} = 0.0905 – 0.0201T + 0.00280T^2$ with an r-squared value of 74.4%, (Severini, 2015, p. 158). This quadratic regression model, shown in Appendix III, does a much better job of capturing the relationship between points scored and time on ice. I was able to grain great insight from this not only in the sense of selecting an appropriate sample to use, but also being careful when analyzing my data and not always sticking with the first model conducted. Therefore, for each of my models, I typically ran multiple analyses and models before selecting the best one and drawing conclusions.

Although the book mainly focused on baseball, football, and basketball, the key takeaway I received from this reading was that it is crucial to keep in mind the variation among time on ice for NHL players. Since some of the top players average over twenty minutes per game, whereas the bottom-line players will only play somewhere between five and ten minutes per game, there is a great advantage for those who play more to then produce more points. There is a strong, positive correlation between time on ice and points per game, therefore, I made sure to only select those players on top-performing lines who will see among the most ice time for their respective team in order to avoid skewing my data. Additionally, it would not make sense to include backup goalies in my sample for the same reason: these goalies will not get the chance to perform to the same level as starter goalies. Therefore, I selected five of the top starting goalies in the league, following the same rules as my other criteria in selected my sample.

The second body of literature related to my topic is titled *Look Who's Talking—Athletes on Twitter: A Case Study* by Ann Pegoraro. This case study investigates the use of social media among athletes, specifically focusing on Twitter. The study acknowledges that social media are a rising force in marketing and have been fully embraced by the sport industry, with teams, leagues, coaches, athletes, and managers establishing presences. Often, athletes use social media to their advantage whether to be to promote themselves or their sponsors. For instance, an athlete may compare themselves to other players to bargain higher salaries, or an athlete may post on behalf of their sponsor for money or to abide by their contract with the company. Primarily these presences have been focused on Twitter, which allows users to post their personal thoughts in 140 characters or less. Athletes, in particular, have engaged in tweeting at a fast pace. This case study investigated the tweets of athletes over a seven-day period in an attempt to find out what and why these athletes were tweeting about. The findings indicate that athletes are talking predominantly about their personal lives and responding to questions or references from fans. The results of this study indicate that Twitter is a powerful tool for increasing fan–athlete interaction.

Although this study does not necessarily study the impact that social media has on performance, this study will have direct relationships to my study. It will prove useful to me as it studies more the intentions behind using social media as an athlete and how it impacts them personally, which in turn, impacts performance. This case studies the value that athletes can place on social media which will be important for me to understand. This review will flow into my thesis because it will give me the understanding I need on the importance of social media in the lives of athletes and their intentions behind using it. Ultimately, this is why I selected this body of literature, as it will give me a strong ground of knowledge to then build upon. In order to construct my thesis to its full potential, I must first understand the basics and the reasoning, purposes, and intentions as each will greatly impact my study.

Studies in the past have also looked at leagues such as the NBA and NFL to see if engagement on Twitter impact performance. Typically, these studies have noticed a relationship among the more vocal players in the league between engagement and performance— the more these players engage in Twitter, typically the worse the perform as a result; however, it will be difficult to replicate this study since NHL players do not engage in social media like these other leagues. Ultimately, this contradicts my study in a sense; although my study will not be focusing on Tweets from the players themselves since there is not enough data. Therefore, I will be focusing on Tweets from the general public and media, which leads me to my take on a study like this. As a result, I believe I will see a more positive relationship present among Tweets from the public and performance as opposed to a more negative relationship among Tweets from players themselves and performance in this body of literature I used. Regardless, knowing this will keep me cautious and allow to compare and contrast my results to the ones I found in this section.

MATERIALS AND METHODS USED

To complete this study, I mirrored the methodologies that I have been exposed to and worked with in my analytics and information systems courses. This involved collecting all appropriate data, cleaning the data, running analyses, building visualizations, obtaining and implementing feedback, and then compiling results and recommendations. I then repeated this process multiple times throughout the duration of completing my thesis.

With the data from the league, specifically the sixty players that I chose to analyze, I ran various summary statistics and tests to clean the data and then visualize it in a clear and concise way. I primary used Python, but also incorporated programs including Excel, Tableau, and SQL to complete my project and run my analyses. Using the data, I focused primarily on summary statistics and correlation analyses to identify any patterns. Then from my results, I attempted to find the causes of these patterns. My main goal of this study was ultimately to explore this topic and the idea of a relationship between Twitter and the NHL, as opposed to running models to find variable importance or accuracy scores, as I was unclear at the beginning as to whether or not there would be any strong indications present in my study.

In order to complete my study, I selected a sample of sixty of the top players from the NHL based on the previous season, spanning every team and position (Appendix IV), consisting of forty forwards, fifteen defensemen, five goalies, and thirty-one teams. Then, I collected season statistics for the eight weeks of my study on each of these players and every team. Next, I collected data from Twitter using key words, accounts, and hashtags about each of the players and teams. I also collected data on the general public's Twitter accounts regarding the NHL or the specific players and teams from the sample. Using all of this data, I then incorporated it into Excel, Python, and SQL in order to complete my intended approach involving collecting all appropriate data, cleaning the data, running analyses, building visualizations, obtaining and implementing feedback, and then compiling results and recommendations. I then repeated this process multiple times throughout the duration of completing my thesis. Finally, I also looked at reports from articles, podcasts, and viewers to see if there is any impact or indications on my study as well.

In order to obtain the appropriate and relevant data to collect, and through working with Professor Kevin Mentzer, we decided that the best way to go about the collection of data that I needed was to pull Twitter data. In collecting Twitter data, I was able to pinpoint various topics or accounts through keywords and handles. Additionally, collecting my data would not only allow me to see these Tweets, but also see the information behind these Tweets, including whether the Tweet is original, a mention, or a retweet, when the Tweet was created, if the user is verified, and how many followers and followings the user account contains. The versatility of collecting data on Twitter is ultimately why I decided to go with this approach for my study.

In order to collect the data, I needed to create a "Twitter listener." Collecting historical Tweets is difficult and costly due to rate limits; furthermore, there are legal issues and restrictions when it comes to collecting past Tweets. The only practical way to grab Twitter data is to write code for a "Twitter listener," which collects Tweets as they are being sent out. This means that this program needs to run constantly to continually collect the tweets needed. The first step in doing this is getting permission through Twitter API. Since there are privacy implications at stake when collecting data from Twitter, I need to apply and get approval on the backend of Twitter in order to continue in the process. This application was relatively painless as almost all accounts are approved when being used for academic purposes. Upon my approval I was able to start collecting my data. This collection then ran through the duration of the period that I was studying.

I worked with Professor Mentzer to write the appropriate Python code necessary to collect my data (Appendix V). Thankfully, I had some prior experience from another class collecting and analyzing Twitter data, but nothing close to this large of a scale. Additionally, I have taken a few classes which focused on Python, so I was relatively comfortable in my abilities to code. Of course, many issues arose, but with the help of my resources I believe that I was able to produce the best result I possibly could. Using a Python package called Tweepy, the process was relatively user-friendly; essentially, every Tweet that met the criteria that I set would be collected and stored in a JSON and CSV file to my desktop, which I could then analyze or upload back to Python or SQL to study and breakdown further. Tweepy was extremely beneficial as it makes it easier to "use the twitter streaming API by handling authentication, connection, creating and destroying the session, reading incoming messages, and partially routing messages (Tweepy)." The code collected the following fields for my dataset:

- Hashtags - The hashtags that appear in the text of a Tweet.
- Tweet ID - The individual ID assigned to every Tweet.
- Created at - The time the Tweet was created.
- Text - The actual Tweet itself.
- Name - The handle the person uses on Twitter (ex. @NHL).
- User - The name the person uses on Twitter besides their handle.
- User ID - The ID given to each user.
- User Location - Where the user is from.
- User Description - The description the user puts in their biography.
- User Followers - The amount of other Twitter users that follow the user.
- User Friends - The amount of people the user follows.
- User Listed - The amount of people the user has added to a list.
- User Created - When the users account was created.
- User Favorites - The number of favorites the user has.
- User Verified - If the user is verified or not.
- User Status - The amount of posts the user has.

The collection of my data was constant throughout the entire season. Since only live tweets can be collected, this code had to be running for about four months total from November to March. There were times when it stopped for a couple of hours, but if that happened, I would turn it back on as soon as I noticed that it had stopped. The annoyances with this were that if there was a period of time where my internet dropped or some other interruption, the code would stop running and need to be restarted. I would check to make sure everything was running smoothly twice per day, and act accordingly if any issues arose. I do not believe that I missed out on much data as the longest my code was not running was probably no more than an hour or two at a time. A more stable and powerful machine or internet aside from my personal laptop and the university internet would allow a more consistent data collection with fewer annoyances such as these.

The criteria that I set to collect my Tweets included each of the thirty-one current NHL team names, as well as their official team handle and hashtag (Appendix VI). Additionally, I included this information for the Seattle Kraken, which will be an expansion team in 2021-2022 NHL season; although, after analyzing my data further, I decided to not include it in my analysis. Additionally, I used key sport accounts such as @NHL, @NHLNetwork, and @TSNHockey, which solely focus on professional hockey in order to avoid skewing my data and by collecting irrelevant Tweets about other sports. Furthermore, these accounts focus on the NHL as a whole, rather than specific teams or players; therefore, my over-arching accounts would not be biased either. Finally, I included the players in my samples accounts and used their last names as keywords as well. Not every player has a Twitter account; furthermore, not many players are vocal on Twitter anyways, as opposed to other leagues such as the NBA or NFL where players are more involved on social media. Regardless, using the players' last names as keywords greatly helped in my data collection. Although, because I was using last names, I had to replace any player in my original sample with last names that are repetitive among the league such as "Smith," as well as last names that are also common words such as "Point." Once I had my criteria set, I was able to start collecting Tweets, which gave me a grand total of around one million Tweets to use. From here, I then organized the Tweets into categories (Appendix VII) such as team, forward, goalie, or defense, to make my analysis by position or topic easier. These categories were then given a key determinant such as "team" to identify which category the Tweet belonged to, and then I appended these identifiers into a new column of the dataset.

The next step was the analysis of the data that I collected. This analysis was mostly done in Python with some tasks being done in Excel and SQL. My main goal was to explore the data at hand to see if there were any interesting trends or patterns involved. From here, I had two main goals: the first was to create a topic analysis of the Tweets I collected. A topic analysis is an analysis of all the topics in the dataset. Topic analyses look at all the Tweets that I collected, and then comes up with a user indicated number of topics that appear in the Tweets that were collected. In my instance, I choose to identify the fifteen most popular topics by returning the most reoccurring string of ten words in the same order. The goal of this is to see what most of the discussion is about in the dataset in order to provide insight as to what people are primarily talking about surrounding the NHL. The NHL as a whole is a large topic but focusing in on these smaller topics was hugely beneficial in understanding what is actually going on in my data and the Twitter world.

The second part of my analysis was the sentiment analysis, which was the most important part of my project. A sentiment analysis looks at the emotion of every tweet in the data set and sees if it is positive or negative by assigning it a score from -1 to 1. Those Tweets with a score closer to 1 are considered more positive, whereas those Tweets with a score closer to -1 are considered more negative. Additionally, if the tweet is scored at 0, then it has neutral sentiment. Depending on the analysis or visualization I was running I could take out any Tweet with a score of exactly 0 or exactly neutral to get a better representation of the scored Tweets. This allowed me to see how people were feeling about the teams and players in my sample throughout the season. My goal was then to compare this to various performance statistics to see if there were any leading or lagging factors in predicting or reflecting performance. For instance, if a certain team or player is generally being talked about very positively, will they then in turn perform at a higher level, or is this just a reflection based on how they have been performing previously. A player performing well is likely going to be talked about positively, but I wanted to see if this were always the case or if any predictions could be made about future performance at a team or individual level. Additionally, using this sentiment analysis, I could then pinpoint different topics or events in the season such as winning or losing streaks, trades, injuries, or other news and see how the Twitter world generally feels about each topic, whether that be positive, negative, or neutral. I broke up my study into different sections, the first being preseason, so anything leading up to the opening night of January 13th, each week of the season (one through eight), and the entire season as a whole from start to finish, this way I could then look at overall sentiment during each of these periods and see how they change week to week or from start to finish. I also collected statistics for each of these players weekly, so I could compare their average sentiment to how they were performing during the current week, that way slumps, time-off, or hot streaks would not skew my data and would be consistent with the current trends at hand.

The third part of my study was the topic analysis, which I used to analyze popular common topics being talked about among my data. From here, I then looked at the social and retweet network to analyze the community, or communities, within this data. This way I could not only see what was being talked about, but by whom and in what manner as well.

Finally, the fourth part of my study focused on the social network and retweet, more specifically, looking into who is talking about or sharing who and why. From this, I was able to look more into which accounts are popular in terms of being retweeted, similarly, those who retweet the most as well. Additionally, with this retweet network, I was also able to look at the sentiment of the Tweets and specific accounts in general to see if this had any impact on my analysis as well.

However, in order to even begin exploring my data, I had to clean it up to be able to work with it and limit any issues down the road. Tweets are messy to work with as they may include punctuation, emojis, basic unnecessary filler words, misspellings, or different capitalization, which can be hugely detrimental since Python is a case-sensitive programming language. In order to eliminate some of these issues I wrote code to remove emojis, turn all text into lowercase, remove punctuation, remove stop words such as “the,” “are,” or “what,” and split up, or lemmatized, the text to have a row for each individual word of the Tweet. This made it easier for differentiating my analyses between players and teams, but furthermore, analyzing official team Twitter accounts versus official team Twitter hashtags. After cleaning up the data, I saved all of it into a new file which only contained formatted Tweets, making it much easier to work with. From here I would then filter the Tweets by date to only contain data from each week of the season and track the patterns among the frequency and usage of each category among players and teams to begin my official analysis. At this point, I was ready to begin exploring my data.

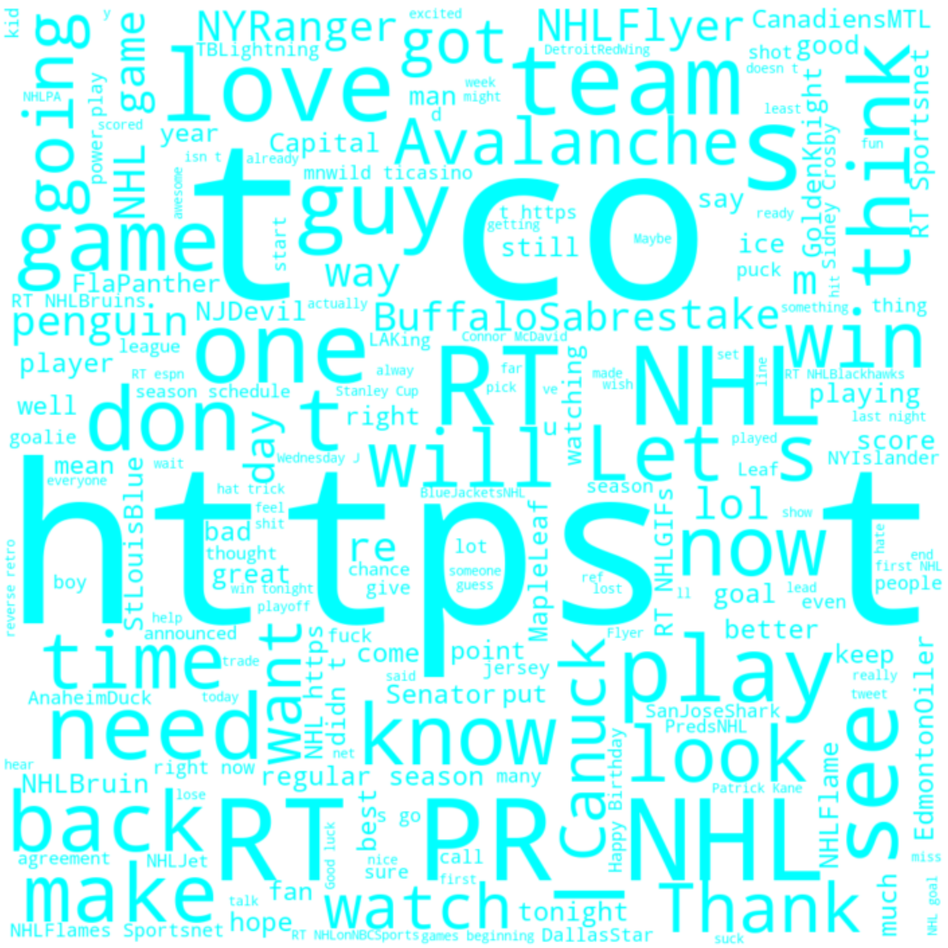
RESULTS AND DISCUSSION

Prior to beginning any analyses on my data, I first had to explore what was at hand in order to see and understand the data I was actually working with.

Exploration Analysis

First, I ran summary statistics to describe my data and found that I had collected 1,045,042 Tweets over the course of the few months that I was working with (Appendix VIII). My original goal was to have about one million Tweets to work with as that would be more than enough, but of course it is always better to be safe and have too many than too few. I was thrilled to have such a strong dataset to work with. Also, I found that there were 288,874 unique accounts contributing to all of these Tweets, which means there is a relatively smaller retweet network present than expected, given that this only accounts for about a quarter of all of the Tweets (Appendix IX). Furthermore, out of all of the Tweets, 32,064 of them were associated with verified Twitter accounts (Appendix X), so this field is primarily discussed among the general public, and less among official accounts. This was expected as typically verified accounts do not retweet often, but since there is quite a large retweet network here, it coincides with the fact that most Tweets are being retweeted (probably from verified accounts) as opposed to original nature Tweets. I wanted to look more into the retweet network to see just how many Tweets are retweeted among the population. I wrote a line of code which then searched through all of the Tweets and counted how many of them contained “RT” indicating that this Tweet is a retweet. After doing so, I found that 273,776 of the Tweets in the dataset were retweets.

As I began to look more into my data, I realized there were a lot of reoccurring words appearing. Many of them made sense such as “NHL” or “hockey,” but others were more surprising. I created a word cloud to visualize the most popular reoccurring words in all of my dataset. Originally, when checking this, words like “the,” “is,” or “what” were among the tops simply because these are very common everyday words used on Twitter. To avoid this from skewing my data I was able to remove basic English “stop words” after importing various packages. From here, all of those filler words were not calculated into my analysis, and the resulting top words (Appendix XI) and coinciding word cloud (Appendix XII) produced very interesting insight:



Of course, words like “NHL” and “RT” appear a lot as previously mentioned, but furthermore, “https” because many people are sharing and retweeting links including videos, news, or highlights of players or teams. However, some of the more insightful findings were the words which contained team names. Clearly, a lot of people are talking about the Dallas Stars, Buffalo Sabres, Montreal Canadiens, New York Ranger, Philadelphia Flyers, and Vancouver Canucks among many other teams, but I then had to find out why.

Beyond the top words, I also wanted to see what the top accounts and hashtags were. As for accounts, the top five most popular accounts (Appendix XIII) were quite surprising as none of them are verified accounts and are typically just fan accounts that Tweet in great volume. On the other hand, the top five hashtags (Appendix XIV) show more expected results including “#NHL” as well as other teams hashtags. This season the NHL implemented two outdoor games to be played beside lake Tahoe. Typically, in the past any outdoor games are just played on a football or baseball stadium, so this was the first of its nature, hence why the hashtag “NHLOutdoors” was so popular in this dataset. Additionally, right before the season commenced in January, Adidas, the current supplier of all jerseys to the NHL, released one new jersey for every team called Reverse Retro Jerseys. These were an immensely popular topic as the jerseys put a new twist on old traditions and team heritage, which explains the other top hashtag, “ReverseRetros.” The other two hashtags surround the Toronto Maple Leafs and the Boston Bruins, both of which have quite large fanbases, and are also rivals as they meet in the playoffs quite frequently. I assume the fanbases account for the popularity of these two hashtags; although, there could be other underlying factors including marketing and advertising schemes which could be skewing the data as well. Similarly, out of the entirety of the retweet network, none of the top five most retweeting accounts are verified, as shown in Appendix XV, and some of these accounts line up with the most popular accounts due to volume. This would make sense since these accounts are quite popular and Tweet very frequently, they must have a larger following and more prone to retweeting. Still, I expected more well-known or reliable accounts to be among the top of this list. I would guess that bigger name, verified accounts report what is already known by the smaller, insider accounts. People love to be the first to know anything, especially with sports, so these smaller inside accounts usually leak information first, whether it be trade rumors, player updates, or team updates, so I would assume that a Twitter user would be more likely to retweet less generally known news and instead retweet the rumored information to be ahead of the game and the rest of the Twitter world.

I wanted to see among my categories: player account, team account, player name, team name, team hashtag, and team location, which was the most popular volume-wise. As shown in Appendix XVI, team hashtags are far and away the most popular component of the Tweets in this dataset. I then broke this down further to see which of the players and teams in my sample are the most popular strictly in regard to volume. As shown in Appendix XVII and Appendix XVIII, Sidney Crosby, Connor McDavid, and Patrick Kane are the most talked about players, presumably as they are often argued to be the best player in the league and are always among the top in terms of performance and general popularity. As for teams, the Colorado Avalanche were the most talked about team, which could be attributed to their strong performance this season as well as a healthy fanbase. Unfortunately, the teams with more than one word in its name, such as the Red Wings or Blue Jackets, did skew my data here as typically people do not Tweet these as one word, yet people will use the word “red” or “blue” when talking about just about anything. Even though this dataset revolves solely around my topic, there is still a high probability that just because the word “red” or “blue” exists in one of these Tweets, that does not mean the Tweet is about the respective team; although, this lapse would not impact the Avalanche given that their name is only one word, so they are still unanimously the most popular in regard to volume alone. However, just because a team or player is frequently Tweeted about, this does not necessarily mean that they are the most popular as they could be Tweeted about frequently in a negative manner. In order to tackle this dilemma, I began my sentiment analysis.

Sentiment Analysis

Following assigning a sentiment score to each of the Tweets in the dataset, I could then understand the manner of these Tweets more. I created a histogram to see the distribution of the sentiment scores (Appendix XIX) and found that most of the Tweets are generally more positive with a mean of about 0.21. I expected to see a great deal of neutral Tweets, which held true, but regardless, there was not a symmetrical distribution among all of the Tweets, which I expected, the data instead is left skewed, implying that there is a large proportion of Tweets above the median, or generally positive. I split the Tweets up into three categories: positive, neutral, and negative, with positive of course representing those Tweets with a score above zero, neutral representing a score equal to zero, and negative representing a score below zero. I expected the vast majority to be neutral Tweets given the previous histogram, but the majority were actually positive, as shown on Appendix XX, so most of the Tweets are just slightly more positive than zero, since even a score as low as 0.0001 would be assigned to the positive category. I then wanted to see what the spread of the data would look like if these neutral Tweets were removed entirely. I ran an analysis to see how the distribution of only positive and negative Tweets compared to the full population (Appendix XXI). This histogram proved that there was a much larger proportion of Tweets with a score above zero than below, with most of these Tweets actually being closer to around 0.5 as opposed to very close to zero. Additionally, I broke down the distribution of the Tweets in each of the three sentiment categories, shown in Appendix XXII, and found that the majority of Tweets were positive, followed by negative, then neutral.

Aside from the Tweets as a whole, I wanted to find which specific words were used the most frequently in positive Tweets versus negative Tweets. I found that these words basically matched each other on opposite sides of the spectrum, implying that these are just extremely common topics, so a great deal of people are talking about them in general, both positively and negatively. Even among Tweets that were entirely in the neutral category, these words stayed just about true to the most popular words overall in regard to volume. Since this idea did not give me any useful insight, I realized I would have to look deeper into my categories.

I then broke this down further to see among which of my categories, including player account, team account, player name, team name, team hashtag, and team location, as shown in Appendix XXIII, were Tweeted about the most positively. All of these categories had a positive overall sentiment score, but the highest was player handles. This was interesting because I expected team hashtags to be the most positive and coincide with the most popularity volume-wise, but this category was actually the third most positive out of the six. Player handles may not be used directly be the player, or owner, themselves, but they are still often Tweeted at or included in various Tweets. Therefore, when people are including player accounts in their Tweets, they are typically the most positive out of all other categories on both the individual player and team levels.

Next, for each specific category, I broke down the Tweets in each classification and ran a sentiment analysis on them. For each of the categories, I looked at the group broken down by sentiment category as well, and I would remove the neutral class from each category to see if this had any impact on the sentiment spread; however, for each category this did not have relatively any impact on the results whatsoever, so for each group, I left all Tweets in regardless of sentiment score or associated category. However, to preface the next section: this is still my exploratory analysis; therefore, I was able to see popularity among sentiment, but I was not yet able to identify the cause. Regardless, from this I was able to make predictions as to why a certain player or team is being generally talked about positively or negatively, but until the weekly breakdown of sentiment in relationship to overall performance, I was unable to actually draw conclusions from my data.

Player Name: As shown in Appendix XXIV, the most positively Tweeted about players overall are Vladimir Tarasenko, Andrei Vasilevskiy, Nikita Kucherov, Sidney Crosby, and Ryan Suter. I found this to be quite interesting as two of these players, Tarasenko and Kucherov, were injured and did not play for just about the entirety of my study. Regardless, these players were spoken about positively, even though I expected this to be the exact opposite as a result of fan or team frustration, especially since all of these players, regardless of whether they are injured or not, are key components of their respective teams.

On the other hand, as shown in Appendix XXV, the players that are most negatively Tweeted about are Mat Barzal, Mark Schiefele, and Mark Giordano. This was intriguing because although this list is similar with popularity in terms of volume of Tweets, these players are actually being spoken about negatively which goes against my original theory regarding these players. As for Barzal, I would assume that he was overall negatively talked about because he was originally in a holdout prior to the season because he was expected to sign a contract extension with the Islanders, but there was talk that they would not be able to afford him and fear that he may not play for quite some time until negotiations were met; therefore, people were most likely speaking negatively in regard to this event. I personally expected to see a player like Jack Eichel on this list as his team, the Buffalo Sabres, are potentially one of the worst teams the league has ever seen, so the frustration of that would most likely be reflected on their captain and star player, Jack Eichel.

Player Account: This category was difficult to draw relevant insight from as there are players among my sample that do not have Twitter accounts; furthermore, among the players that do have accounts, many of them do not engage in them frequently. Regardless, these accounts are used by the general public quite frequently, so this analysis was still worthwhile.

As shown in Appendix XXVI, the most positively used player account would be Alexander Ovechkin, followed by Brendan Gallagher, Steven Stamkos, Auston Matthews, and Anze Kopitar. I did not find this very surprising as these are very respected players among the league, all of which wear a captaincy or alternate captaincy letter for their respective team.

On the other hand, as shown in Appendix XXVII, the most negatively used player accounts are Brad Marchand, Elias Petterson, and Mitchell Marner. As for Marner, who overlaps here among both positive and negative categories, as with many other topics, this is most likely due to just volume alone. However, I would think that Marner's negativity could also be due to his recently signed contract which is widely considered to be far too large for him. Elias Petterson on the other hand, was off to a horrible start this season following a very promising rookie and sophomore season, which was mostly likely why he was often Tweeted about negatively due to his performance. Additionally, Brad Marchand was expected to be on this list as he is notorious across the league for being a "rat" or a pest— a chippy player who loves to get under his opponent's skin. As a result, he has a reputation of being hated by just about everyone aside from Boston Bruin fans where he plays, causing him to then be spoken about negatively in turn.

Team Name: As for the team name category, shown in Appendix XXVIII, I decided not to use this analysis to make any predictions in my study because of the bias that multi-word team names create. This analysis resulted in "wings," "blues," and "golden" being the most positively Tweeted about; however, these do not necessarily identify their respective team and could be attributed to other Tweets. It can be assumed that teams like the Blackhawks and Avalanche are truly positively spoken about, but this is not reliable enough to make any legitimate predictions; therefore, I depended more on team accounts and hashtags are these are one word and officially licensed.

Team Account: As shown in Appendix XXIX, the most positively Tweeted about team accounts are the Calgary Flames, New York Rangers, and Nashville Predators. Each of these teams have strong fanbases, perform well, and a strong rival team as well which could be the underlying cause of this. Especially since each team is limited to their conference this year in their regular season opponents, teams typically play their rival eight times throughout the year, which only adds more fuel to the rivalry fire. The Calgary Flames and Edmonton Oilers arguably have one of the best rivalries in the NHL right now in what is called "The Battle of Alberta."

On the other hand, the most negatively Tweeted about teams include the Anaheim Ducks, Winnipeg Jets, and New Jersey Devils. Similar to the Sabres, the Ducks were off to a horrendous start to the season, barely competing with any team they were playing. They have since then been able to turn things around, but as my study focuses on preseason and the first half of the season, I would assume that there was a lot of frustration during this time and not enough of a cool-down period to smooth out the negativity. The Jets coincides with their player, Mark Schiefele, being on the negative player name list, but as for them as well as the Devils, I am currently not sure why they are being talked about so negatively.

Team Hashtag: As shown in Appendix XXX, the most positively used team hashtags are the Minnesota Wild, Toronto Maple Leafs, and Anaheim Ducks. This was the most surprising category to me since I could not determine a cause for the Wild and Maple Leafs to be this positively used aside from popularity or volume alone, but the most surprising aspect was that the Ducks team account was among the most negatively used, yet their team hashtag is among the most positive.

On the other hand, the Arizona Coyotes, Winnipeg Jets, and Buffalo Sabres are among the most negatively used team hashtags. I could not determine any event or factor that would lead to the Coyotes being at the bottom of this list besides a relatively weak fanbase compared to other NHL teams. The Jets for whatever reason keep reappearing on the negative side of the spectrum, so that will be something to keep an eye on going forward, and as previously mentioned, the Sabres are currently the league's worst performing team, and the players and fans are extremely frustrated with their lackluster performance and poor management.

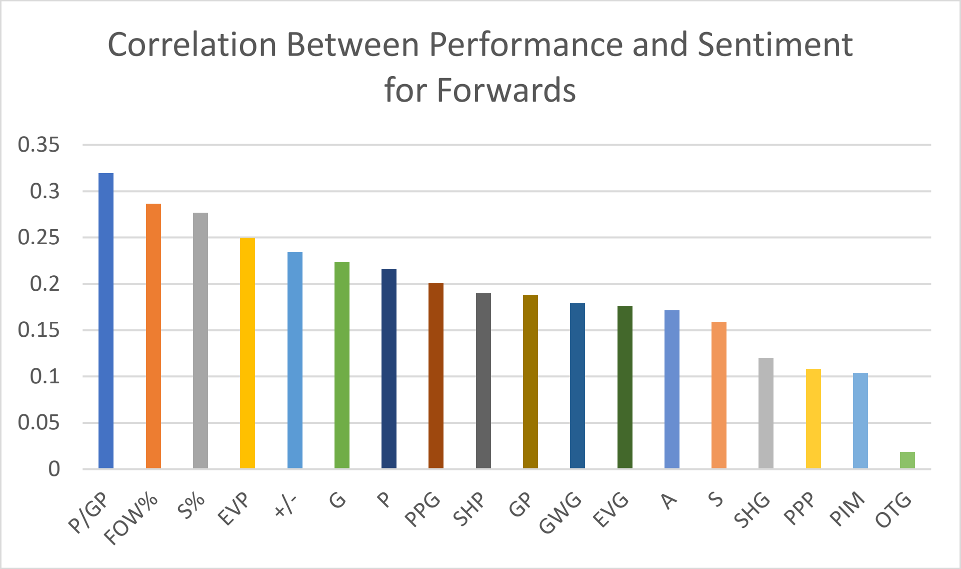
Team Location: Similar to team name, I decided not to utilize this analysis to make predictions. There is too much risk associated with this category as well since location can be talking about much more than just a city's hockey team.

Now, understanding the over-arching sentiment breakdown of the Tweets and categories, I was confident in the "what" of my study, but now I just needed to account for the "why." In breaking down the Tweets by week and comparing sentiment to performance, I believed I would be able to see if sentiment is predictive or just reflective of performance. More specifically, I wanted to see if Player X in my sample is talked about positively, would they in turn perform better or worse even, or just if this positive talk was simply a result of prior performance, whether that be weak or strong. Therefore, even if the relationship was simply reflective, I wanted to see if the correlation between sentiment was strictly positive, which I expected since if a player performs poorly, they will probably be talked about poorly and vice versa, or if there were other instances in my data. There are obviously many other factors that play into sentiment scores aside from performance, such as trades, contracts, personal life, injuries, and more, which will be discussed later on in my topic analysis.

I broke down my data into weeks running from Wednesday to the following Tuesday of each week of the first half of the season, accounting for eight weeks total. The main point of this was to see if there would be any key events whether it be game, and performance related or personal life related that could account for the weekly sentiment scores. As for my sentiment analysis, I focused solely on the season as a whole, rather than each individual week, as that will be discussed in my topic analysis. The sentiment analysis consisted of five main categories: forwards, defensemen, goalies, official team accounts, and official team hashtags. I kept forwards, defensemen, and goalies in separate categories because they have different statistics. Obviously, goalies have far different game statistics than forwards and defensemen including save percentage, goals against average, and wins; however, forwards and defensemen have different statistic lines as well. For example, defensemen would not have faceoff win percentage as only forwards take faceoffs. Furthermore, I did not include powerplay statistics such as powerplay or shorthanded points, for example, as forwards have a far greater advantage here. Forwards typically comprise a team’s powerplay unit and therefore have a far greater advantage when it comes to potentially producing points. Similarly, on the penalty kill, teams are defensive minded and do not typically get the opportunity to produce points; however, it does happen from time to time, but regardless, when this does occur, forwards are almost always the ones responsible for these points. As for the differentiation between team account and team hashtag, there are not different statistic lines for these as the team statistics remain constant; however, as seen earlier, the Twitter world uses these two identifiers very differently, so I wanted to see how this would impact overall sentiment. As a result, I stuck with these five categories in my sentiment analysis.

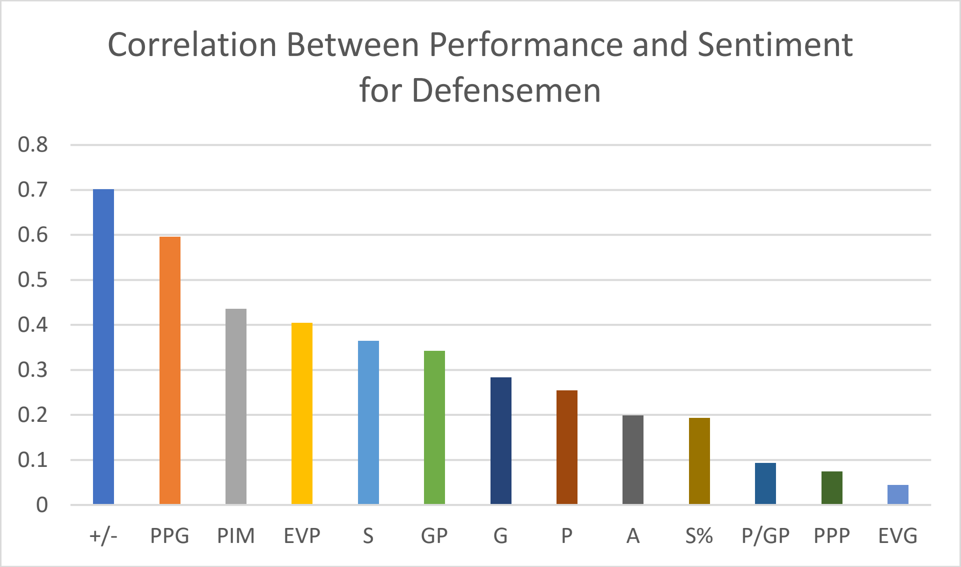
For each of these categories, I ran a correlation matrix between average sentiment and each statistic line to see which aspects had the strongest correlation to sentiment. Therefore, for instance, I would be able to see if say points had a very strong, positive correlation to sentiment, then higher sentiment could potentially account for higher point production. Each player category (forwards, defensemen, and goalies) included all of those players from my sample as a whole; therefore, the categories consist of forty forwards for the forwards category, fifteen defensemen for the defensemen category, and five goalies for the goalie category. Additionally, for each mean correlation score, I converted these to absolute value to make visualizations easier to comprehend; therefore, a correlation of 0.4 and -0.39 would be next to each other on a chart as -0.39 would be represented as 0.39. I figured I just wanted to see if there were any strong correlations and then determine if those were positive or negative afterwards to make my analysis easier to understand. The true correlations will be shown in the appendix, which will be noted, respectively. To begin, I started with forwards group:

Forwards:



The charted correlations show that points per game played has the strongest correlation to sentiment, suggesting that these two are closely related. This was originally what I expected to see; however, the relationship is actually negative, as shown in Appendix XXXI, implying that a higher overall sentiment score is related to fewer points per game that a player will record. I was extremely surprised to find this; although, the correlation is still relatively weak. Regardless, it is the strongest recorded correlation among all of the statistic lines, and the only category above 30%. As a result, I would not necessarily conclude that this relationship exists since the score is too low to conclude anything too promising.

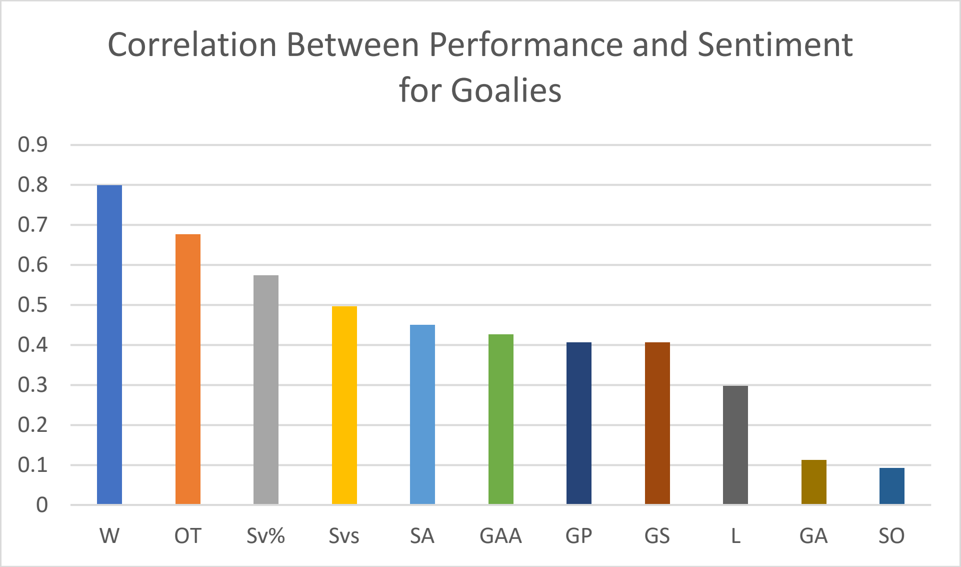
Defensemen:



The defensemen category showed far more promising results, as shown above. The strongest correlation existed between average sentiment and plus-minus (+/-) at just over 0.7. +/- is a statistic that represents a player's overall impact on the game, as it signifies the difference between their team's total scoring versus their opponent's when the player is on the ice. Therefore, if a player is on the ice when their team scores their plus-minus would be +1, but if the other team score while they are on the ice, their plus-minus would be -1. If they are not on the ice when a goal is scored, their plus-minus does not change. Ultimately, this statistic represents goal differential on an individual level.

The correlation between plus-minus and mean sentiment score is notable strong and positive, as shown in Appendix XXXII, which suggests that a higher overall sentiment has a strong relationship to individual goal differential, or plus-minus. Although correlation does not imply causation, there is still a strong relationship present here, which is extremely promising in my original prediction that sentiment relates to performance.

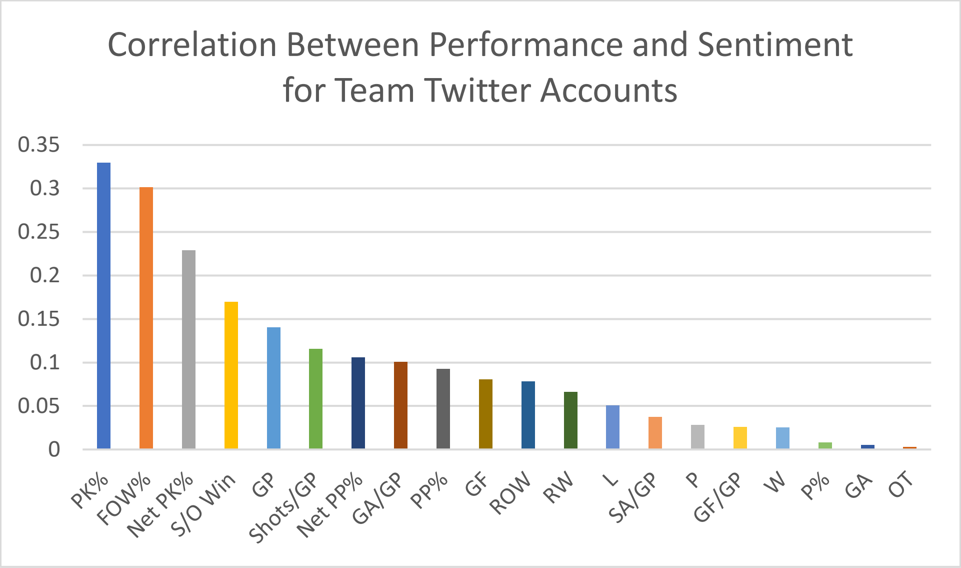
Goalies:



As for goalies, I also found extremely promising results. The chart above shows that there is a strong correlation between average sentiment and wins at about 0.8. This correlation is strong and positive, as shown in Appendix XXXIII, which implies that a positive relationship exists between sentiment and wins, more specifically, higher average sentiment relates to higher win totals as a goalie. This also coincides with my original prediction that higher sentiment would lead to higher performance. Surprisingly, the correlation between mean sentiment and overtime wins (OT) is relatively strong too; however, this relationship is actually negative, implying that there is a relatively strong relationship, more specifically, a higher average sentiment typically leads to fewer overtime wins. Although this partially contradicts the previous conclusion about wins, overtime only occurs if a team does not win in regulation, which would then be categorized as a win. There are far more factors that play into overtime periods, and it actually reiterates the fact that a goalie may be more likely to win in regulation and not have to play in an overtime period when those goalies have a higher sentiment.

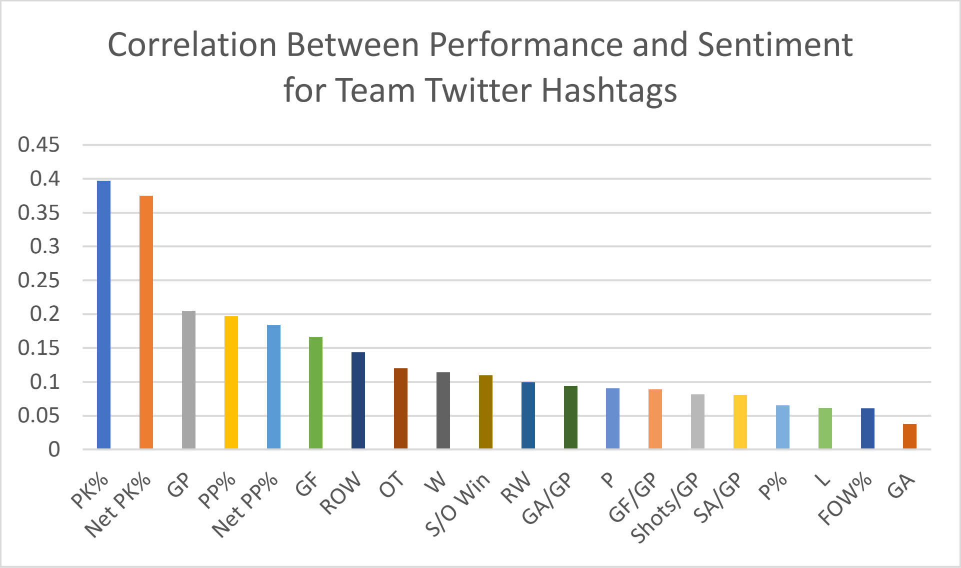
As for other, more individualized statistics such as save percentage and total saves, these are also relatively corelated to mean sentiment at about 0.57 and 0.5, respectively. Both of these correlations are positive, so they do not contradict each other, and both imply that there is a relationship, although relatively weak, between save statistics for a goalie and overall sentiment score.

Official Team Accounts:



As shown in the graphic above, there were not any strong correlations among statistic lines and mean sentiment, as was found with the forwards group. The strongest correlation would be between penalty kill percentage (PK%) and sentiment, which represents the proportion of times a team is able to survive being short-handed due to a penalty on their team without being scored upon. This correlation is actually negative, as shown in Appendix XXXIV, at about -0.33, which would imply that there is a negative relationship, more specifically a team with a higher average sentiment score may be less successful on the penalty kill. Regardless, this correlation is not strong enough to provide any truly meaningful results or insights.

Official Team Hashtag:



As previously mentioned, I kept team accounts and hashtags separate as they are used differently on Twitter, therefore, I expected to find different results; although, as shown above, the results were quite similar to the official team account category. Likewise, the strongest correlation here to mean sentiment is with penalty kill percentage at about 0.4; however, this relationship is actually positive now (Appendix XXXV), implying the opposite relationship between penalty kill and average sentiment exists for official team hashtag. I found this quite interesting that they would be just about the exact opposite for this team statistic; however, the correlation is still not very strong to conclude this relationship is stable. Regardless, having penalty kill at opposite sides of the correlation spectrum is quite interesting given that they are both statistics for the same teams and the only difference here is that the negative correlation exists when using the official team account in a Tweet and the positive correlation exists when using the official team hashtag in a Tweet. I also attempted to compare both team accounts and hashtags to fanbase sizes to see if this Twitter data could be related to number of fans; although, I did not find anything significant enough to conclude for this section.

Following this analysis, it became obvious that correlation and relationships do exist in my data. The next step was to then see if there were any leading factors among the relationships between statistics (both player and team) and mean sentiment, or if these were all lagging factors, meaning that sentiment on Twitter is simply reflective or performance rather than predictive, or even both reflective and predictive. In order to do this, I now had to use my weekly breakdown and sentiment scores to my advantage on a team and individual level to see if I could account for some of the pits or peaks in average sentiment.

For this analysis, I used the three most fluctuating teams and players, so these teams and players had the highest standard deviation among all of their weekly sentiment scores (Appendix XXXVI). First, for the team analysis, I used the Coyotes, Stars, and Canucks average sentiment over all eight weeks (Appendix XXXVII). From here, I broke down each of these teams into separate graphs (Appendix XXXVIII) and compared their average sentiment scores to win percentage for the week. For instance, if one of these teams played four games in the fourth week, their win percentage would be 0.75. I charted the team's sentiment against win percentages (Appendix XXXIX), and for each of these three teams, I found that typically sentiment was reflective of win percentage; therefore, a higher win percentage typically led to a higher sentiment score in the same or following week, whereas a lower win percentage tended to lead toward a lower sentiment score in the same or following week.

I wanted to replicate this on an individual level, so I looked at the three most fluctuating players, so the three players with the highest standard deviation among their sentiment scores as well (Appendix XL). I decided to compare these players' scores to their respective team's win percentage as well because it was quite difficult to find any relationship among these players and any of the statistical categories. It appeared that teams carried more weight in accounting for the fluctuations of sentiment, as a similar pattern was present when looking at the individual level (Appendix XLI). Likewise, sentiment appeared to follow similar trendline patterns to the player's respective team's win percentage; however, this time, the relationship appeared to go more hand-in-hand with one another, as opposed to the sentiment being reflective. This is probably because the sentiment is still grouped by week as opposed to a day-by-day level, so this would probably still be a reflective factor. I assume that I found very similar results for the teams and players for this analysis because teams carry more weight and it is very uncommon for Tweets to be conducted that talk about a certain player without mentioning their team to some degree, whereas it is common for Tweets to be conducted about teams and have no mention of specific players.

It was becoming clear that sentiment was solely a reflective aspect of performance and events as opposed to being predictive. I thought there was a chance that if Twitter users hyped up a certain team or player they would perform better as a result or vice versa; however, on Twitter, it is obvious that these users are simply reacting to how teams and players perform following games or other events. Now that I knew this, I wanted to see what exactly was being talked about in my data. I obviously knew this data focused on hockey, but being such a large topic, I wanted to get more specific than that and find actual topics that are being talked about within this larger theme.

Topic Analysis

For the next major part of my analysis, I wanted to see what topics were primarily being talked about in common throughout my dataset. I broke this portion up into three different subcategories: topic analysis, event analysis, and social and retweet network analysis. More specifically, the topic analysis would look at the various commonalities in discussions among the Twitter users in my data to see which topics were being talked about the most, the event analysis would look directly at events which occurred in the league during my study and analyze those and the reaction from the Twitter world, and the social and retweet network analysis would specifically focus on the retweets in my data and see if any communities existed in my data, meaning groups all talking about the same thing, or if everyone was just talking about the NHL or hockey in gender. Obviously, I knew I was going to find a great deal of topics, but I wanted to focus on the most popular topics found. In order to do this, I built a Latent Dirichlet Allocation (LDA) model using a Python package called Gensim, which broke up the Tweets into categories based on commonalities of the series of words within the Tweet. From here, I was able to find various topics which contained similar discussions or key words. I was able to produce an interactive chart which listed each of the topics as well as the key words in each, Appendix XLII. Once I had these topics and words, I had to interpret them in order to predict what is actually being talked about here. For instance, the LDA model would return a list of words that would not compile directly into a legible sentence, instead, it would resemble broken English, which I would then have to interpret to the best of my abilities. I stuck with analyzing the top four topics, as following these, the topics seemed to get repetitive or irrelevant to my topic. These four were both relevant and relatively unique, so I decided to only include these in my analysis.

Topic Breakdown

The first, most popular, topic seemed to revolve primarily around the NHL season coming back and puck drop. This would make sense as there was a lot of uncertainty among the league this season and whether or not the NHL would actually be able to undergo a season, and if so, to what capacity. Therefore, a great number of users were talking about this uncertainty on Twitter, especially once the start date and league specifics and COVID related guidelines were announced. Additionally, puck drop ties into the season opener, indicating puck drop to start the season, but more so, to start each individual game as well. Typically, when anticipating a game, fans, analysts, coaches, players, and more will refer to "puck drop" as the mark of the beginning of the contest. This is often talked about greatly on Twitter given that games are highly anticipated and frequent, so it gives users an increased opportunity to Tweet about this topic.

The second most popular topic involved the creation of the NHL Reverse Retro jerseys by Adidas. As previously mentioned, these jerseys were new to the league this year, and it was the first time that every team across the league implemented a new jersey into their rotation all at once. These were hugely popular and discussed, especially on Twitter since Adidas used their social media presence to their advantage and advertised the new jerseys on platforms such as Twitter. Users then reacted to the news once they were released, as well as any other time a team would wear them during a game. The outdoor games solely featured these new jerseys from each of the four teams that participated that weekend, so there was a steady discussion about these, as opposed to a spike at the beginning when they were released, and then having the conversation die out.

The third most popular topic involved competition, winning, and optimism. Typically, this seemed to just be a reflection on various games and performance, as users would react to games using key phrases such as "winning" or "competition." Additionally, this could be used when anticipating a game as well in a similar manner. I was not entirely surprised that "optimism" was so popular, since I think of sports fans typically being optimistic about their team's status and ranking. I was surprised, however, that it was listed within the third most popular topic.

Finally, the fourth most popular topic involved the purchasing of the Reverse Retro jerseys. Whereas the second topic focused on the creation of these jerseys, this topic focused on the buying aspect. This would mean that the discussion involved in this topic was primarily from fans buying these jerseys and then talking about their recent purchase on Twitter; this could also involve users Tweeting about wanting to buy one of the jerseys as well. Given that the second and fourth topics were identified as two different topics from one another, this emphasizes how large of an impact the jerseys had on the Twitter world. I believe this ultimately goes hand in hand with strong advertising and marketing done by Adidas in the creation and selling of these jerseys; although, I will discuss later on that marketing to a large group would be difficult in this community. Additionally, the conversation was constant among my study since these jerseys came out right around the beginning of my data collection and continued to be worn by teams and purchased by fans throughout the entirety of my study.

Event Analysis

I then wanted to look at key events that happened during the season and see how Twitter reacted to these. I chose events that revolved around my sample of players and teams and were anticipated and could both be talked about leading up to the event as well as afterwards to prevent bias in this analysis. The four topics I chose on the individual player level were Sidney Crosby’s 1000th game played, Vladimir Tarasenko returning from his injury, Artemi Panarin taking a leave of absence from the league, and Patrick Kane’s 1000th game. As shown in Appendix XLIII, you can see how the Twitter world reacted to each of these events in terms of overall sentiment.

During the week of Crosby’s 1000th game, there was a large spike in overall sentiment, showing that people were often talking in a positive manner about Crosby and his major accomplishment. As for Kane and the same milestone, people did not typically talk about him as positively and he actually experienced a decrease in sentiment leading up to this event, and my study concluded before I could capture the sentiment following this event.

As for Panarin, who had to take a leave of absence following allegations made about him after he spoke out about the Russian government. He returned back home to Russia to deal with these claims, and the Twitter world typically increased in positivity when talking about him right before and after his decision; however, my study concluded before I could see how Twitter users reacted upon his return.

Finally, Tarasenko saw an increase before and an even stronger one upon his return, which would indicate that Twitter users were quite pleased with him being back in the lineup. This event is quite interesting because the upward and downward trends almost mirror his injury progress identically, since he was hopeful to return around week four and then had a setback, causing a decline in overall sentiment, and finally this large increase once he announced his return timetable.

Following this, I decided to replicate a similar analysis, but this time on a team level. For this I focused on the outdoor Lake Tahoe games which were new to the league. This came up a great deal throughout my study between the game itself, the teams and players, and the Reverse Retro jerseys that they would be wearing. This took place on the weekend of week six of my study. As shown in Appendix XLIV, the Avalanche, Bruins, and Flyers each saw a spike in overall sentiment to at least some degree, which would indicate that Twitter users were overall positive about this new experience. The Bruins experienced the largest increase following their win against the Flyers. I found it to be surprising that the large increase came after the game, however, as opposed to before when the league was advertising and attracting media to the event. I would think that the event and discussion around the outdoor games as a whole started high and only got higher as it came closer to that weekend, but soon fell after these games were over as they were not as successful as projected between poor weather and ice conditions, but on a team level, it would make more sense that the sentiment would increase as a result of a win against a rival team.

Social and Retweet Network Analysis

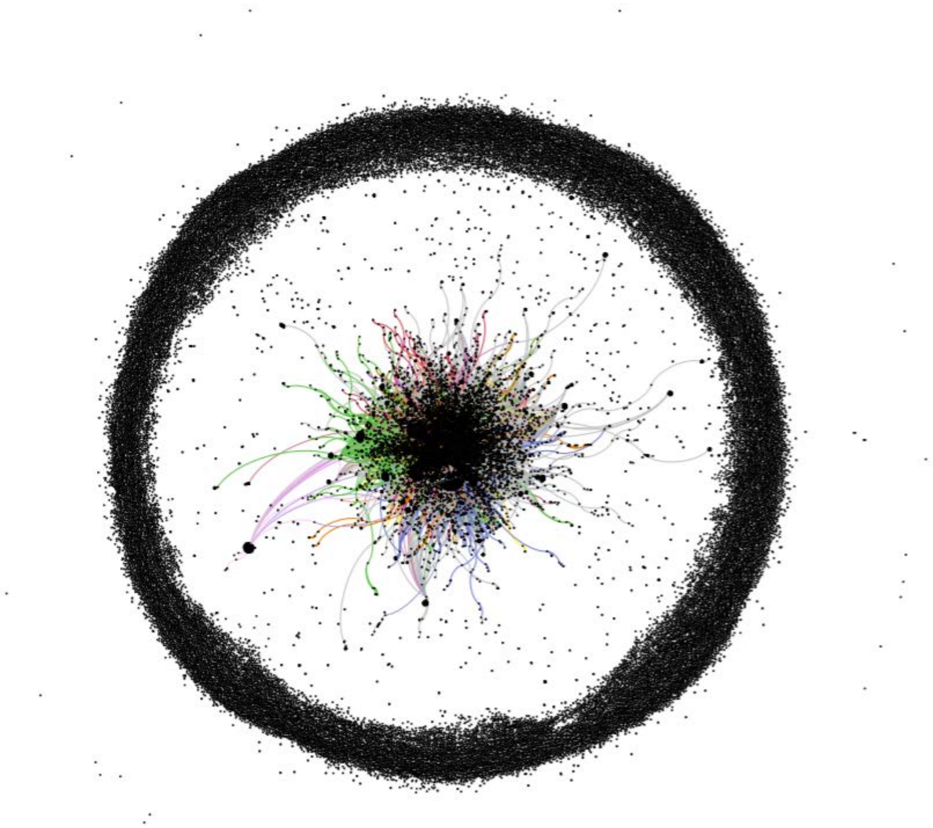
Next, I wanted to look specifically at the retweet network within my data, so the first thing I did was create a new data frame consisting only of Tweets which contained “RT” in the text indicating that this was a retweet instead of an original Tweet. As previously mentioned, the retweet network only made up about one quarter of my data, which was surprisingly low, so I wanted to see if I could explain why that was. First, I looked at the retweet network in regard to my topics from my LDA model and found that there was a heavy leading account in multiple topic nodes. As shown in Appendix XLV, the leading account in the first topic, or node, is @ivagonefishing, who is an avid Colorado Avalanche fan and is constantly retweeting specifically relating to Tweets within the first topic, which contained the discussion surrounding the NHL season being pack and puck drop. The other accounts representing the nodes similarly are just fan accounts that retweet in large volumes. Only one of these accounts is actually verified, shown in orange, on the chart in the appendix, and almost all of these accounts Tweet positively on average with the exception of one account. These results were not hugely important to my analysis since they were just fan accounts retweeting in large volumes, but it helped me gain initial insight on this social network at hand.

As a whole, I found that the retweet network similarly reflected the same sentiment breakdown overall as the entire dataset. Since these retweets are just recycled original Tweets that are also in my dataset, this would make sense, as it is essentially a sample group of my population. As a result, the spread of the sentiment of these retweets is relatively normal but slightly left skewed, indicating mostly positive retweets, and even more left skewed when the neutral retweets were removed. Both of these are shown in Appendix XLVI and Appendix XLVII respectively, but mirror very similar results to the data at large.

I then wanted to account for the community as a whole and see who is talking about or retweeting who to see if NHL on Twitter can be broken down into subcommunities talking about various topics with one another. For instance, in an analysis involving politics, the results here would show that there are two main subcommunities present, representing democrats and republicans, and these subcommunities would typically not interact with one another as they would simply discuss among their own group. I wanted to see if this could be replicated among the NHL, so perhaps, different teams or divisions could make up subcommunities. Additionally, I expected there to be a subcommunity that did not revolve around the NHL. Given the pandemic and the recent presidential election, everyone seemed to be discussing these topics, regardless of the main objective in Tweeting. Politics tend to come up in some shape or form in analyses like mine, so I was expected to see something of that nature.

For this, I used a Python package called Gephy which was able to tell me who was retweeting who and identify these subcommunities or nodes. This also gave me more in-depth retweet network statistics, including the accounts which retweet the most (Appendix XLVIII), which is represented by a statistic line called “out degree” in Gephy, as well as which account are retweeted the most (Appendix XLIX), which is represented by “in degree.” I was surprised to see that the Bruins retweet the most, which would imply that they have the most active social media network. I was surprised to find that the Bruins were ahead of large sports networks such as NBC Sports and Fox Sports, as well as the main NHL account as well. As for the most retweeted accounts, NHL was at the top, which was exactly what I was expecting to find (Appendix L). Some of the other accounts in the top ten of each respective list here, specifically the ones that are not verified, can be presumed to be robot accounts. These accounts make other fake accounts to retweet what is posted on their main account in mass quantities, to get their name out more and get their account to show up as much as possible on Twitter. Therefore, the accounts like @Mcguiretipping and @Dakotadamus seemed to be surprising finds at first but turned out to not provide any meaningful insights to my study. However, as expected I found that typically those accounts which retweet more typically do not get retweeted as much and vice versa as the more official or credible accounts typically get retweeted and do not retweet themselves, whereas the fans or analysts tend to be the ones retweeting these higher up accounts and do not typically find themselves being the ones who are retweeted.

Next, I wanted to look specifically at the community breakdown, if there was one. In this analysis, I found that within this network, the average connections were quite low at 0.6. This means that on average, these accounts either retweet or get retweeted 0.6 times total. There was an overwhelming number of accounts with a connection value of zero, which caused my results to be slightly skewed here. Although, the weighted connection average was about 3.1, which is the total number of times retweeting plus the number of times being retweeted. Similarly, these results are still skewed by those accounts with a connection score of zero, but it makes sense that this would be higher as it accounts for the sum of both sides of the retweet spectrum. Using my results from Gephy, I got the following visualization here and in Appendix LI as a result:



The results from Gephy ultimately showed that there is only one main community present here, which is represented by the circle as a whole. Each color represents various nodes, within the community, but these nodes are ultimately discussing the same thing, which would be the NHL or hockey to some degree. The outer ring represents those accounts which have neither retweeted nor been retweeted more than once. These accounts are either relatively inactive, or just throw out Tweets without being active in responding or reacting to the rest of Twitter with this topic. The other nodes, or colors, do not make up much of the community at large, since the largest node only makes up about six percent of the community (Appendix LII) or about twenty-four percent of the top ten nodes (Appendix LIII). Although, there is only one community here, this is still a very interesting find because I did not expect everyone in my dataset to be discussing the same topic amongst each other. I fully expected there to be subcommunities not interacting, whether it be due to rivalries, which was a stretch, but more likely because discussions about politics or COVID-19 I thought would surely arise. However, there were not enough of these Tweets relating to these topics and still being relevant to the NHL or my study in some way in order to be collected in the first place by my Twitter listener in order to be significant enough to show up as a separate community. As a result, even though this community is talking about the same things, there is not a healthy back and forth discussion taking place here given that there is only one community, and the accounts within it seem to be parallel playing with one another. People appear to just be throwing Tweets out there as there is not a lot of sharing existing. Therefore, marketers should not pay to get their messages out there using this niche on Twitter, as the largest node only covers about six percent of the community. More specifically, no one has a dominant presence when it comes to talking about hockey on Twitter.

Using Gephy, I was also able to find which accounts had the top prestige score as well (Appendix LIV). These accounts represent the nodes that are most connected to the high retweet accounts. However, given that there are not large communities detected here, this score does not have as much weight as it would where multiple communities are present, like in politics. Overall, at first these findings from Gephy seemed to be underwhelming but finding that there is only one community with no dominant presence or existence of back-and-forth discussion among this field of Twitter was quite interesting. This would imply that when it comes to hockey on Twitter, users, for the most part, just Tweet and are not heard as they do not engage with one another. Other topics would typically not find results similar to this as they would probably find at least some healthy discussion going on amongst the community or communities.

CONCLUSION

For this portion, I decided to breakdown my conclusion into each of the key parts of my study.

Exploration Analysis

The top words in hashtags had to do directly with the NHL, but more specifically, those teams which were present here typically have the strongest fanbases (Appendix LV) and were dominant in this category due to volume of fans, which in turn, increased volume of Tweets about the respective team. Other top words or hashtags discussed the Lake Tahoe outdoor games and the Adidas Reverse Retro jerseys, as these were both new and talked about greatly in this community. Overall, the average sentiment score for all of the Tweets collected was about 0.21, so slightly positive. Positive Tweets made up about 47.6% of the data, so the spread of sentiment score was mostly positive and left skewed as a result, especially once the neutral category was removed. The top players and teams in terms of average sentiment could not necessarily be explained for the most part; however, specific weekly events and performance could explain the week-to-week sentiment fluctuations.

Sentiment Analysis

Once I ran correlation matrices each week for every statistic category against average sentiment for teams and players, I found that +/-, or goal differential, had the strongest correlation for defensemen at about 0.7, indicating that a higher sentiment score could be correlated to a higher goal differential. Furthermore, I found that the strongest correlation for goalies was between average sentiment and wins at about 0.8, indicating that a higher sentiment score could be correlated to a higher win total. Both of these statistics are often discussed when talking about performance for the respective positions. Unfortunately, I did not find any other correlations strong enough for forwards or teams as a whole that I considered to be worth including; although, win percentage often accounted for fluctuations in team and player average sentiment on a weekly basis.

Twitter is typically reflective of performance or other events, which accounts for the lagging factors and other visualizations discussed. When a team or player performs well, they usually experience an increase in sentiment as a result or vice-versa, which was exactly what I predicted.

Additionally, various events were reacted to accordingly depending on if fans were happy or upset about a given event. For example, fans were excited about the new jerseys or the outdoor games, so teams and this topic saw an increase in average sentiment. On the other hand, teams had to cancel games throughout the season, so typically when this would occur, the respective teams or players would see a decrease in sentiment, and then this would level out, or often increase once the team resumed to play. Ultimately, the Tweets in my data focused on current events and performance and reflected these, as opposed to predicting future events or statistics, whether good or bad.

Topic Analysis

The four most popular topics discussed the season being back, puck drop, competition, prior year, and reverse retro jerseys. All of these ultimately made sense to my study, but I was surprised that there was not anything listed about politics or COVID-19; however, some of the more obscure words that appeared were typically just robot accounts getting in the way because they Tweet or retweet in mass quantities automatically. In each of these topic nodes, the dominating account almost always had a positive average sentiment score, and were almost never a verified account, aside from one account for each of these statements.

Social Network Analysis

Since only about a quarter of the data was accounted for by retweets, the vast majority of the accounts neither retweet nor get retweeted; therefore, there is not a lot of back-and-forth sharing. There is only one community talking about the NHL here and there is no dominant presence surrounding this topic. Typically, users are just throwing Tweets out there and not interacting or sharing with one another. Furthermore, the largest subcommunity only makes up about six percent of the community at large, so there are really no marketing opportunities worthwhile for exploiting here. The top retweeting accounts show which teams or accounts have the most prominent social media presence, whereas the top retweeted accounts show which teams or accounts are the most trusted or used.

Going Forward

There were obviously limitations with my study, mostly because of COVID, which impacted scheduling, games, teams and players, fans, and more. Replicating this study for a normal season from start to finish would be quite interesting. Additionally, it would be interesting to look at other forms of social media aside from just Twitter, as there could be differences or further insight on this topic or idea if replicated across other sports or leagues. Going forward, it would be nice if this concept could be used for more predictive modeling, perhaps predicting playoff picture, player and team awards, draft picks, and more; however, in order to do this, there would have to be data collected for much longer, most likely including the entirety of the off-season as well. Unfortunately, since I did not really find any predictive factors present, it would be quite difficult to do any predictive modeling here, especially since this is a team sport, and it would be difficult to pinpoint specific statistical categories to use as factors here. An individual sport like golf for instance would probably be much easier to model. On the other hand, it could be possible to use team standings to conduct models on this data, but again, this would probably also have to be a normal season since this season some teams would be lower in the standings just because they have played fewer games due to COVID cancellations. A discrepancy exists here as a result, since a team may have only played five games and be compared to a team that has played ten or more. However, in a normal season, standings would be more consistent and hopefully an analysis could be done regarding how sentiment plays into performance, or make predictions regarding future team standings, which would tie into playoff projections as well.

It would also be interesting to see this replicated more across other teams and leagues, especially since players from other leagues engage in Twitter much more than NHL players do. I think similar results could be found, but it would most likely be easier to make predictions as well as opposed to just reactions. Therefore, we might actually be able to find that we can hype a person or team up and they will perform better as a result, as I had originally hoped I would find. I think more can be done in terms of specific game breakdown as well, so looking at sentiment prior to, during, and following games to see how sentiment is impacted in comparison to game performance. It would be interesting to see if certain players experience a temporary, or possibly extended boost in sentiment on average, whereas other players may not experience the same boost, and then try to attack the question of why. A study including this would potentially be able to predict what a player's sentiment will be on Twitter after a game is played and if this is only because of points scored or if there are other factors involved as well.

Ultimately, my main goal in this study surrounded exploration as opposed to modeling, and I think this was a smart move because it would be quite difficult to do any sort of predictive modeling here. I think my study does, however, provide great insight for someone who would want to expand upon this in terms of attempting to model and predict this or another topic. Regardless, this study also provides strong insight on exploring this or a similar topic, as I believe I covered all of the bases I set out to tackle, but I would love to see what more can be done with this work, as it can be widely replicated and hugely beneficial to teams, analysts, fans, and more.

ACKNOWLEDGMENTS

I would like to thank all of the faculty, staff, and peers who have helped me along the way with this study and in completing the Bryant University Honors Program. Specifically, I would like to thank my advisor, Kevin Mentzer, for his willingness to work with me and immense support and assistance throughout the entirety of the process. As well as my editorial reviewer, Dr. Suhong Li, who has also been a great resource for me throughout my time at Bryant. Finally, my academic advisor, Dustin Lesperance, who has always been an unbelievable help with anything and everything, even beyond the Honors Program.

I am beyond thankful for the support I have received throughout my time at Bryant, especially in the midst of working on my thesis. This experience has been nothing but rewarding, and I am so thankful for having the opportunity to share my work. Thank you.

APPENDIX

Appendix I: Severini Sample

Name	GP	P	TOI	P/GP
Daniel Winnik	84	23	16.7	0.27381
Cody Hodgson	83	41	13.81667	0.493976
Steven Stamkos	82	97	22.01667	1.182927
Ryan Getzlaf	82	57	21.6	0.695122
Eric Staal	82	70	21.55	0.853659
Zach Parise	82	69	21.48333	0.841463
Anze Kopitar	82	76	21.33333	0.926829
Dany Heatley	82	53	20.95	0.646341
Joe Pavelski	82	61	20.61667	0.743902
Jarome Iginla	82	67	20.6	0.817073
John Tavares	82	81	20.56667	0.987805
Patrick Marleau	82	64	20.48333	0.780488
Joe Thornton	82	77	20.46667	0.939024
Brad Richards	82	66	20.26667	0.804878
Patrick Kane	82	66	20.2	0.804878
Dustin Brown	82	54	20.16667	0.658537
Phil Kessel	82	82	20.05	1
David Backes	82	54	20	0.658537
Henrik Zetterberg	82	69	19.83333	0.841463
Loui Eriksson	82	71	19.76667	0.865854
Jason Pominville	82	73	19.68333	0.890244
Andrew Ladd	82	50	19.56667	0.609756
Marian Gaborik	82	76	19.51667	0.926829
Matt Moulson	82	69	19.3	0.841463
Tomas Fleischmann	82	61	19.1	0.743902
Henrik Sedin	82	81	19.08333	0.987805
Rick Nash	82	59	19.08333	0.719512

Ryan Smyth	82	46	19.08333	0.560976
Kyle Brodziak	82	44	19.06667	0.536585
Olli Jokinen	82	61	18.96667	0.743902
Derek Stepan	82	51	18.95	0.621951
Dainius Zubrus	82	44	18.68333	0.536585
Ray Whitney	82	77	18.65	0.939024
Gabriel Landeskog	82	52	18.61667	0.634146
Erik Cole	82	61	18.53333	0.743902
Brooks Laich	82	41	18.5	0.5
Bobby Ryan	82	57	18.35	0.695122
Chris Kunitz	82	61	18.31667	0.743902
Patrik Berglund	82	38	17.96667	0.463415
Vinny Prospal	82	55	17.88333	0.670732
Teemu Selanne	82	66	17.88333	0.804878
Scott Hartnell	82	67	17.78333	0.817073
Frans Nielsen	82	47	17.45	0.573171
Brandon Sutter	82	32	17.4	0.390244
Michael Ryder	82	62	17.38333	0.756098
Antoine Vermette	82	37	17.2	0.45122
Troy Brouwer	82	33	17.18333	0.402439
Justin Williams	82	59	17.15	0.719512
Lauri Korpikoski	82	37	17.13333	0.45122
Pascal Dupuis	82	59	16.93333	0.719512
Wayne Simmonds	82	49	15.91667	0.597561
Petr Sykora	82	44	15.9	0.536585
Matt Cooke	82	38	15.68333	0.463415
Colin Greening	82	37	15.58333	0.45122
Brian Boyle	82	26	15.23333	0.317073
Jannik Hansen	82	39	14.9	0.47561
Alexei Ponikarovsky	82	33	14.75	0.402439
Chris Kelly	82	39	14.73333	0.47561

Andrew Cogliano	82	26	14.7	0.317073
Nick Foligno	82	47	14.65	0.573171
Jason Chimera	82	39	14.43333	0.47561
Vernon Fiddler	82	21	13.98333	0.256098
Darroll Powe	82	13	13.98333	0.158537
Mikkel Boedker	82	24	13.63333	0.292683
Brandon Prust	82	17	11.95	0.207317
Craig Adams	82	18	11.28333	0.219512
Maxim Lapierre	82	19	11.23333	0.231707
Tomas Plekanec	81	52	20.75	0.641975
Marian Hossa	81	77	19.96667	0.950617
Patrik Elias	81	78	19.85	0.962963
Shawn Horcoff	81	34	19.58333	0.419753
Ryan O'Reilly	81	55	19.53333	0.679012
Patrice Bergeron	81	64	18.58333	0.790123
David Desharnais	81	60	18.4	0.740741
Valtteri Filppula	81	66	18.26667	0.814815
Milan Lucic	81	61	17.03333	0.753086
Milan Hejduk	81	37	17.01667	0.45679
Tyler Seguin	81	67	16.93333	0.82716
Teddy Purcell	81	65	16.13333	0.802469
Maxime Talbot	81	34	16	0.419753
Jiri Hudler	81	50	15.66667	0.617284
Erik Condra	81	25	14.16667	0.308642
Zack Smith	81	26	14.06667	0.320988
Tom Kostopoulos	81	12	12.31667	0.148148
Justin Abdelkader	81	22	12.31667	0.271605
Jamal Mayers	81	15	9.8	0.185185
Kyle Clifford	81	12	9.4	0.148148
Shawn Thornton	81	13	9.183333	0.160494
Corey Perry	80	60	21.38333	0.75

Stephen Weiss	80	57	20.51667	0.7125
Jason Spezza	80	84	19.91667	1.05
T.J. Oshie	80	54	19.53333	0.675
Derek Roy	80	44	19.31667	0.55
James Neal	80	81	19.13333	1.0125
Blake Wheeler	80	64	19.08333	0.8
P.A. Parenteau	80	67	18.65	0.8375
Logan Couture	80	65	18.56667	0.8125
Alex Burrows	80	52	18.46667	0.65
Drew Stafford	80	50	17.65	0.625
Tomas Kopecky	80	32	17.26667	0.4
Marcus Johansson	80	46	16.8	0.575
David Clarkson	80	46	16.36667	0.575
Josh Bailey	80	32	15.21667	0.4
Samuel Pahlsson	80	17	14.76667	0.2125
Jay McClement	80	17	13.75	0.2125
Drew Miller	80	25	12.86667	0.3125
Matt Martin	80	14	12.15	0.175
Scott Nichol	80	8	9.316667	0.1
Shane Doan	79	50	19.6	0.632911
Paul Stastny	79	53	18.83333	0.670886
David Krejci	79	62	18.41667	0.78481
Max Pacioretty	79	65	18.26667	0.822785
Jussi Jokinen	79	46	17.66667	0.582278
Kyle Okposo	79	45	17.06667	0.56962
Matt Read	79	47	17.06667	0.594937
Dominic Moore	79	25	15.53333	0.316456
Chris Stewart	79	30	15.43333	0.379747
Ryan Jones	79	33	15.43333	0.417722
Artem Anisimov	79	36	15.4	0.455696
Steve Sullivan	79	48	15.36667	0.607595

Lars Eller	79	28	15.31667	0.35443
Jiri Tlusty	79	36	14.9	0.455696
Viktor Stalberg	79	43	14.06667	0.544304
Shawn Matthias	79	24	13.81667	0.303797
Raffi Torres	79	26	11.36667	0.329114
Tim Brent	79	24	10.88333	0.303797
Alex Ovechkin	78	65	19.8	0.833333
David Legwand	78	53	18.51667	0.679487
Jordan Eberle	78	76	17.6	0.974359
Thomas Vanek	78	61	16.93333	0.782051
Jarret Stoll	78	21	16.68333	0.269231
Jakub Voracek	78	49	16.28333	0.628205
Michael Grabner	78	32	15.55	0.410256
Jim Slater	78	21	14.76667	0.269231
Eric Belanger	78	16	14.73333	0.205128
Andrew Brunette	78	27	13.55	0.346154
Jamie McGinn	78	37	13.45	0.474359
Tanner Glass	78	16	13.41667	0.205128
Gregory Campbell	78	16	12.8	0.205128
Manny Malhotra	78	18	12.35	0.230769
Matt Hendricks	78	9	12.11667	0.115385
Ilya Kovalchuk	77	83	24.43333	1.077922
Martin St. Louis	77	74	22.63333	0.961039
Claude Giroux	77	93	21.55	1.207792
Ryan Kesler	77	49	20.1	0.636364
Milan Michalek	77	60	19.55	0.779221
Radim Vrbata	77	62	18.65	0.805195
R.J. Umberger	77	40	18.18333	0.519481
Johan Franzen	77	56	17.7	0.727273
Alexander Semin	77	54	16.78333	0.701299
Brandon Dubinsky	77	34	16.26667	0.441558

Nick Spaling	77	22	15.71667	0.285714
Kyle Wellwood	77	47	14.95	0.61039
Derek Dorsett	77	20	14.7	0.25974
Nick Johnson	77	26	14.45	0.337662
Sean Couturier	77	27	14.13333	0.350649
Jordin Tootoo	77	30	13.15	0.38961
Anthony Stewart	77	20	8.116667	0.25974
Ryan Callahan	76	54	21.03333	0.710526
Ryane Clowe	76	45	17.86667	0.592105
Rene Bourque	76	24	17.83333	0.315789
Brad Marchand	76	55	17.61667	0.723684
Alex Burmistrov	76	28	16.66667	0.368421
Dave Bolland	76	37	16.5	0.486842
Patric Hornqvist	76	43	15.33333	0.565789
David Steckel	76	13	12.83333	0.171053
Torrey Mitchell	76	19	12.43333	0.25
Andrew Desjardins	76	17	9.583333	0.223684
Evgeni Malkin	75	109	21.01667	1.453333
Daniel Alfredsson	75	59	18.95	0.786667
Sam Gagner	75	47	17.18333	0.626667
Sergei Kostitsyn	75	43	16.46667	0.573333
Danny Cleary	75	33	15.98333	0.44
Steve Downie	75	41	15.93333	0.546667
Boyd Gordon	75	23	15.93333	0.306667
Colton Gillies	75	8	10.08333	0.106667
Tim Jackman	75	7	9.116667	0.093333
Cody McLeod	75	11	7.2	0.146667
Bryan Little	74	46	20.21667	0.621622
Mike Ribeiro	74	63	20.05	0.851351
Patrick Sharp	74	69	19.9	0.932432
Mike Richards	74	44	18.88333	0.594595

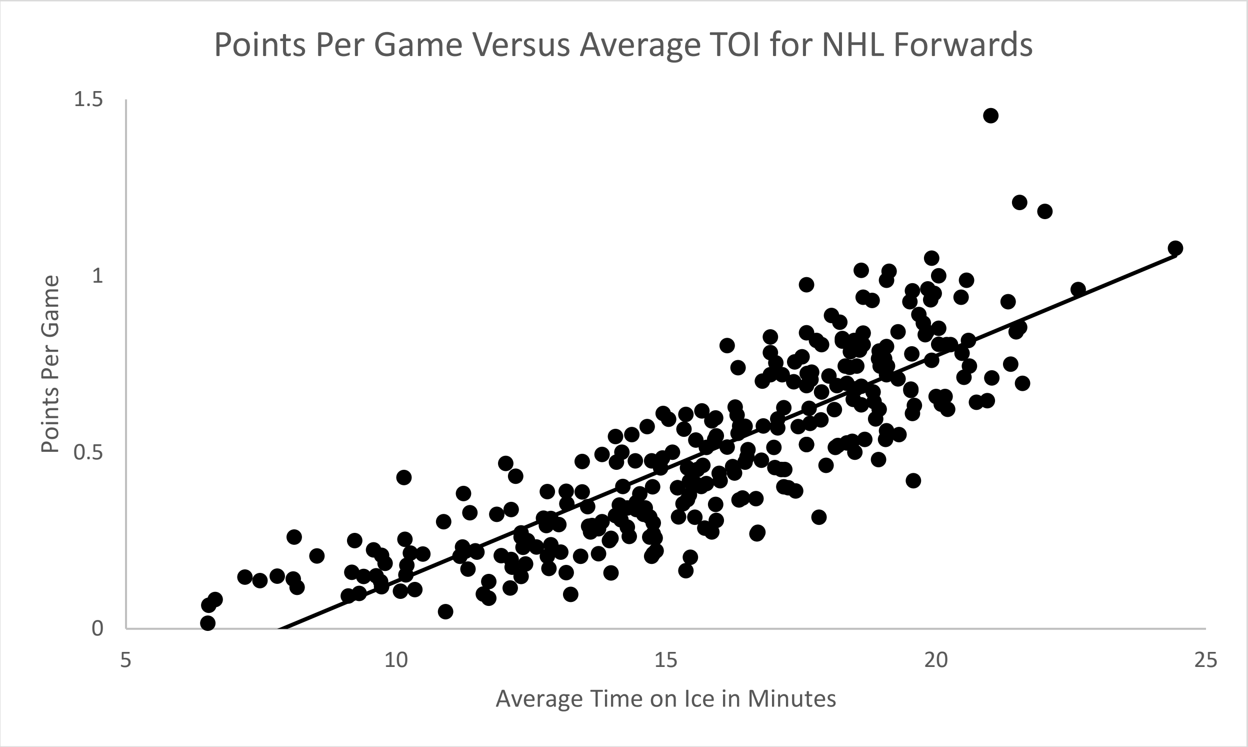
Steve Ott	74	39	18.35	0.527027
Adam Henrique	74	51	18.16667	0.689189
Saku Koivu	74	38	18.13333	0.513514
Mikhail Grabovski	74	51	17.6	0.689189
Evander Kane	74	57	17.51667	0.77027
Cal Clutterbuck	74	27	16.35	0.364865
Derick Brassard	74	41	16.33333	0.554054
Blake Comeau	74	15	15.45	0.202703
Tom Pyatt	74	19	14.8	0.256757
Eric Nystrom	74	21	13.75	0.283784
Benoit Pouliot	74	32	12.21667	0.432432
Tomas Holmstrom	74	24	11.86667	0.324324
Matt Cullen	73	35	18.93333	0.479452
Tyler Bozak	73	47	18.85	0.643836
Jaromir Jagr	73	54	16.33333	0.739726
Clarke MacArthur	73	43	15.85	0.589041
Vladimir Sobotka	73	20	15.85	0.273973
Daymond Langkow	73	30	15.75	0.410959
Patrick Dwyer	73	12	15.36667	0.164384
Radek Dvorak	73	21	14.28333	0.287671
Ruslan Fedotenko	73	20	13.6	0.273973
Joel Ward	73	18	12.43333	0.246575
Taylor Pyatt	73	19	12.31667	0.260274
Matt Halischuk	73	28	11.25	0.383562
Mike Fisher	72	51	19.3	0.708333
Daniel Sedin	72	67	18.81667	0.930556
Tuomo Ruutu	72	34	16.46667	0.472222
David Jones	72	37	15.75	0.513889
Andrei Kostitsyn	72	36	15.11667	0.5
Craig Smith	72	36	14.18333	0.5
Jason Arnott	72	34	14.08333	0.472222

Mike Knuble	72	18	13.95	0.25
Trevor Lewis	72	7	13.23333	0.097222
Chris Neil	72	28	12.8	0.388889
Ryan Carter	72	8	10.35	0.111111
Chris Thorburn	72	11	10.18333	0.152778
Tom Wandell	72	15	9.733333	0.208333
Kris Versteeg	71	54	19.91667	0.760563
Martin Erat	71	58	18.48333	0.816901
Jamie Benn	71	63	18.06667	0.887324
Chris Higgins	71	43	16.31667	0.605634
Ville Leino	71	25	15.91667	0.352113
Todd Bertuzzi	71	38	15.55	0.535211
Marcus Kruger	71	26	15.4	0.366197
Niklas Hagman	71	23	14.58333	0.323944
Bryan Bickell	71	24	12.13333	0.338028
Cory Emmerton	71	10	8.1	0.140845
Pavel Datsyuk	70	67	19.56667	0.957143
Danny Briere	70	49	17.36667	0.7
Tim Connolly	70	36	17	0.514286
Nikolai Kulemin	70	28	15.21667	0.4
Paul Gaustad	70	21	14.76667	0.3
Jamie Langenbrunner	70	24	14.61667	0.342857
Brian Rolston	70	24	14.28333	0.342857
Matt Beleskey	70	15	10.26667	0.214286
Ales Hemsky	69	36	17.6	0.521739
Devin Setoguchi	69	36	17.6	0.521739
Nik Antropov	69	35	16.51667	0.507246
T.J. Galiardi	69	15	13.05	0.217391
Jeff Halpern	69	16	12.6	0.231884
Jerred Smithson	69	6	11.71667	0.086957
Daniel Paille	69	15	11.5	0.217391

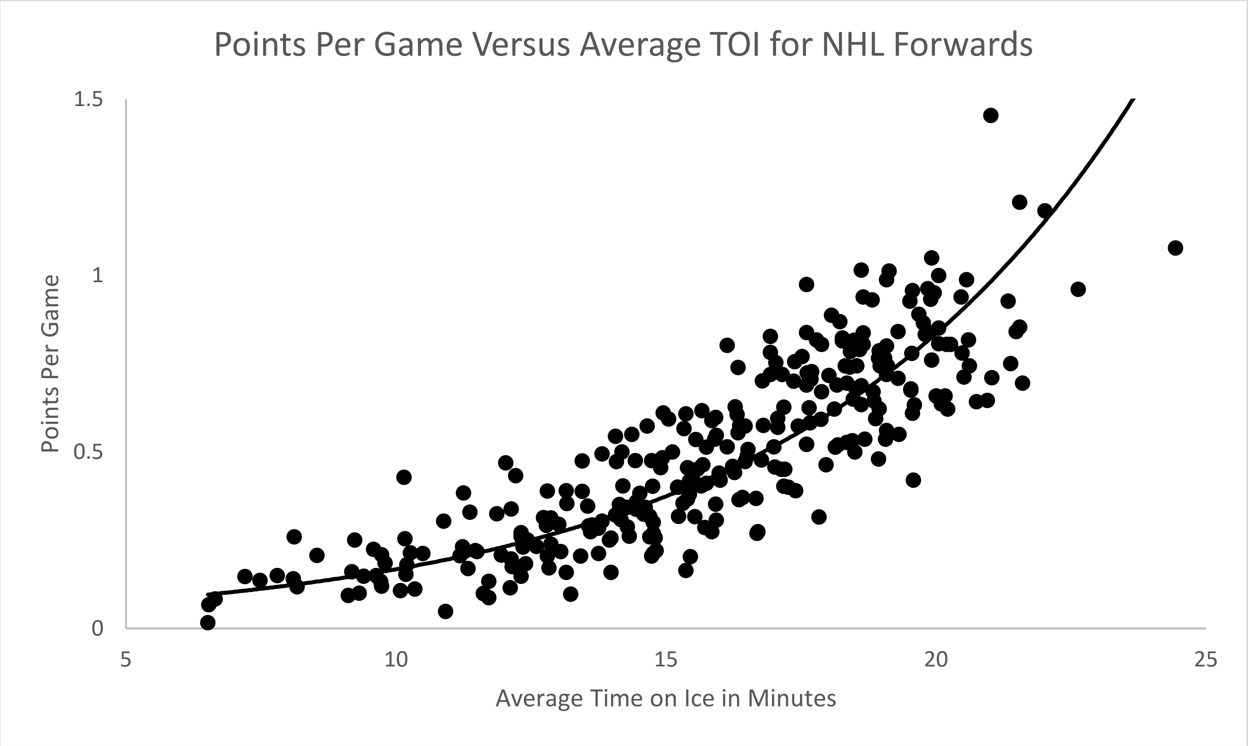
Ryan Malone	68	48	17.68333	0.705882
Colin Wilson	68	35	16.13333	0.514706
Nate Thompson	68	15	14.81667	0.220588
Darren Helm	68	26	14.51667	0.382353
Brett Connolly	68	15	11.46667	0.220588
Joe Vitale	68	14	11.18333	0.205882
Dale Weise	68	8	8.166667	0.117647
Curtis Glencross	67	48	18.01667	0.716418
Chad LaRose	67	32	16.76667	0.477612
Michal Handzus	67	24	14.45	0.358209
Joey Crabb	67	26	13.45	0.38806
Ryan Johansen	67	21	12.73333	0.313433
Colin Fraser	67	8	9.733333	0.119403
Brad Winchester	67	10	7.8	0.149254
Joffrey Lupul	66	67	18.61667	1.015152
Mike Cammalleri	66	41	18.11667	0.621212
Derek MacKenzie	66	14	10.5	0.212121
Zac Rinaldo	66	9	7.483333	0.136364
Dustin Penner	65	17	14.31667	0.261538
Brad Boyes	65	23	13.16667	0.353846
Adam Burish	65	19	12.78333	0.292308
Kaspars Daugavins	65	11	11.33333	0.169231
Alex Tanguay	64	49	19.05	0.765625
Vincent Lecavalier	64	49	18.93333	0.765625
Jeff Skinner	64	44	18.61667	0.6875
Martin Hanzal	64	34	18.45	0.53125
Carl Hagelin	64	38	15.05	0.59375
Mathieu Perreault	64	30	12.03333	0.46875
Arron Asham	64	16	9.233333	0.25
Patrick Kaleta	63	10	13.15	0.15873
Michael Frolik	63	15	12.86667	0.238095

John Mitchell	63	16	10.16667	0.253968
Tim Stapleton	63	27	10.15	0.428571
Ben Eager	63	13	8.533333	0.206349
Jordan Staal	62	50	20.05	0.806452
Ryan Nugent-Hopkins	62	52	17.6	0.83871
Sean Bergenheim	62	23	16.41667	0.370968
Mark Letestu	62	25	15.65	0.403226
David Booth	62	30	14.93333	0.483871
Nathan Gerbe	62	25	14.2	0.403226
Matthew Lombardi	62	18	13.56667	0.290323
Jay Pandolfo	62	3	10.91667	0.048387
Brad Staubitz	62	1	6.516667	0.016129
Taylor Hall	61	53	18.21667	0.868852
Lee Stempniak	61	28	16.23333	0.459016
Matt Stajan	61	18	13.01667	0.295082
Mathieu Darche	61	12	12.13333	0.196721
Marty Reasoner	61	6	11.61667	0.098361
Roman Horak	61	11	10.2	0.180328
Tyler Kennedy	60	33	14.36667	0.55
Mike Santorelli	60	11	12.4	0.183333
Andreas Nodl	60	8	11.71667	0.133333
Matt Ellis	60	8	9.716667	0.133333
Lennart Petrell	60	9	9.633333	0.15
Mike Rupp	60	5	6.65	0.083333
Ryan Reaves	60	4	6.533333	0.066667

Appendix II: Severini Linear Regression Model



Appendix III: Severini Quadratic Regression Model



Appendix IV: Team and Player Sample

Forwards	**Team**
Sebastian Aho	CAR
Aleksander Barkov	FLA
Mathew Barzal	NYI
Patrice Bergeron	BOS
Sean Couturier	PHI
Sidney Crosby	PIT
Leon Draisaitl	EDM
Jack Eichel	BUF
Brendan Gallagher	MTL
Johnny Gaudreau	CGY
Claude Giroux	PHI
Nico Hischier	NJ
Jonathan Huberdeau	FLA
Patrick Kane	CHI
Clayton Keller	ARI
Travis Konecny	PHI
Anze Kopitar	LA
Nikita Kucherov	TB
Dylan Larkin	DET
Nathan MacKinnon	COL
Evgeni Malkin	PIT
Brad Marchand	BOS
Mitchell Marner	TOR
Auston Matthews	TOR
Connor McDavid	EDM
Alex Ovechkin	WHS
Max Pacioretty	VGK
Artemi Panarin	NYR
David Pastrnak	BOS
Elias Pettersson	VAN

Mikko Rantanen	COL
Mark Scheifele	WPG
Tyler Seguin	DAL
Steven Stamkos	TB
Andrei Svechnikov	CAR
Vladimir Tarasenko	STL
Tomas Tatar	MTL
John Tavares	TOR
Blake Wheeler	WPG
Mika Zibanejad	NYR

Defensemen	**Team**
Brent Burns	SJ
John Carlson	WSH
Thomas Chabot	OTT
Mark Giordano	CGY
Victor Hedman	TB
Miro Heiskanen	DAL
Seth Jones	CBJ
Roman Josi	NSH
Erik Karlsson	SJ
Torey Krug	STL
Cale Makar	COL
Alex Pietrangelo	VGK
Jaccob Slavin	CAR
Ryan Suter	MIN
Zach Werenski	CBJ

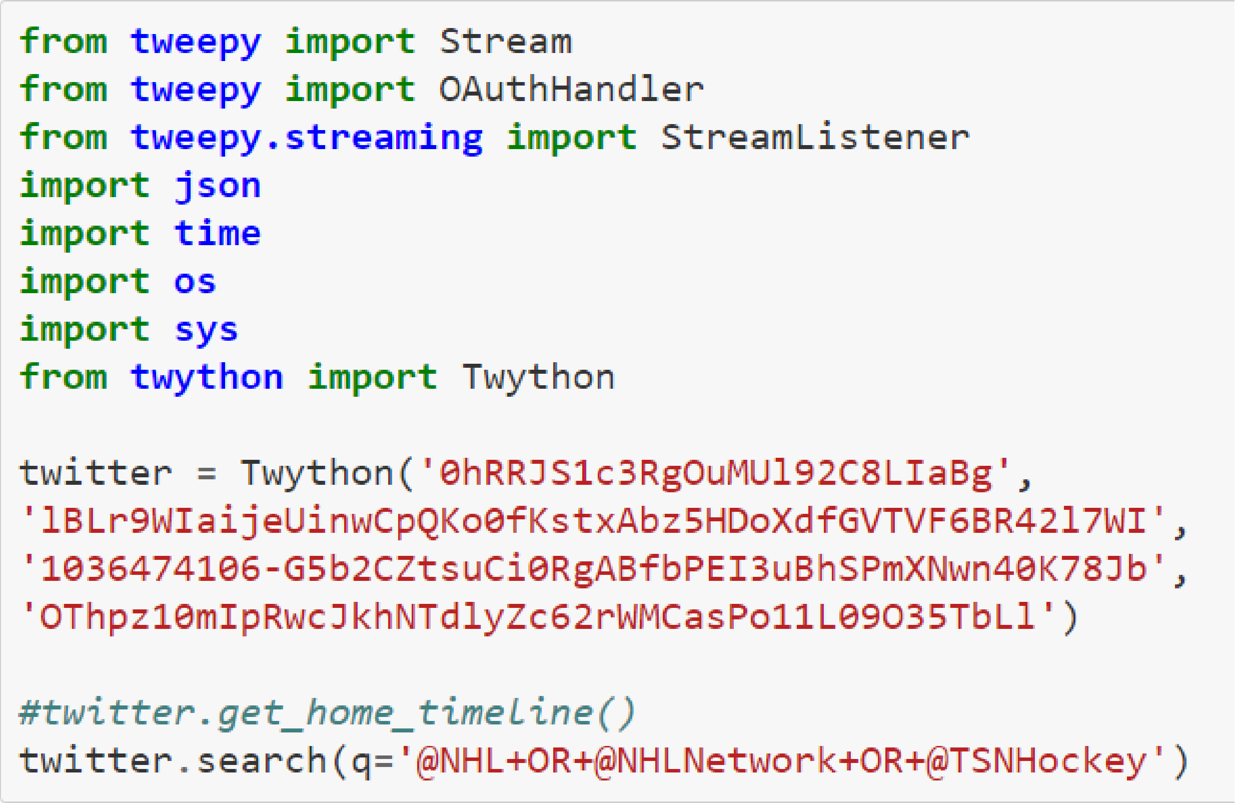
Goalies	**Team**
John Gibson	ANA

Connor Hallebuyck	WPG
Tuukka Rask	BOS
Semyon Varlamov	NYI
Andrei Vasilevskiy	TB

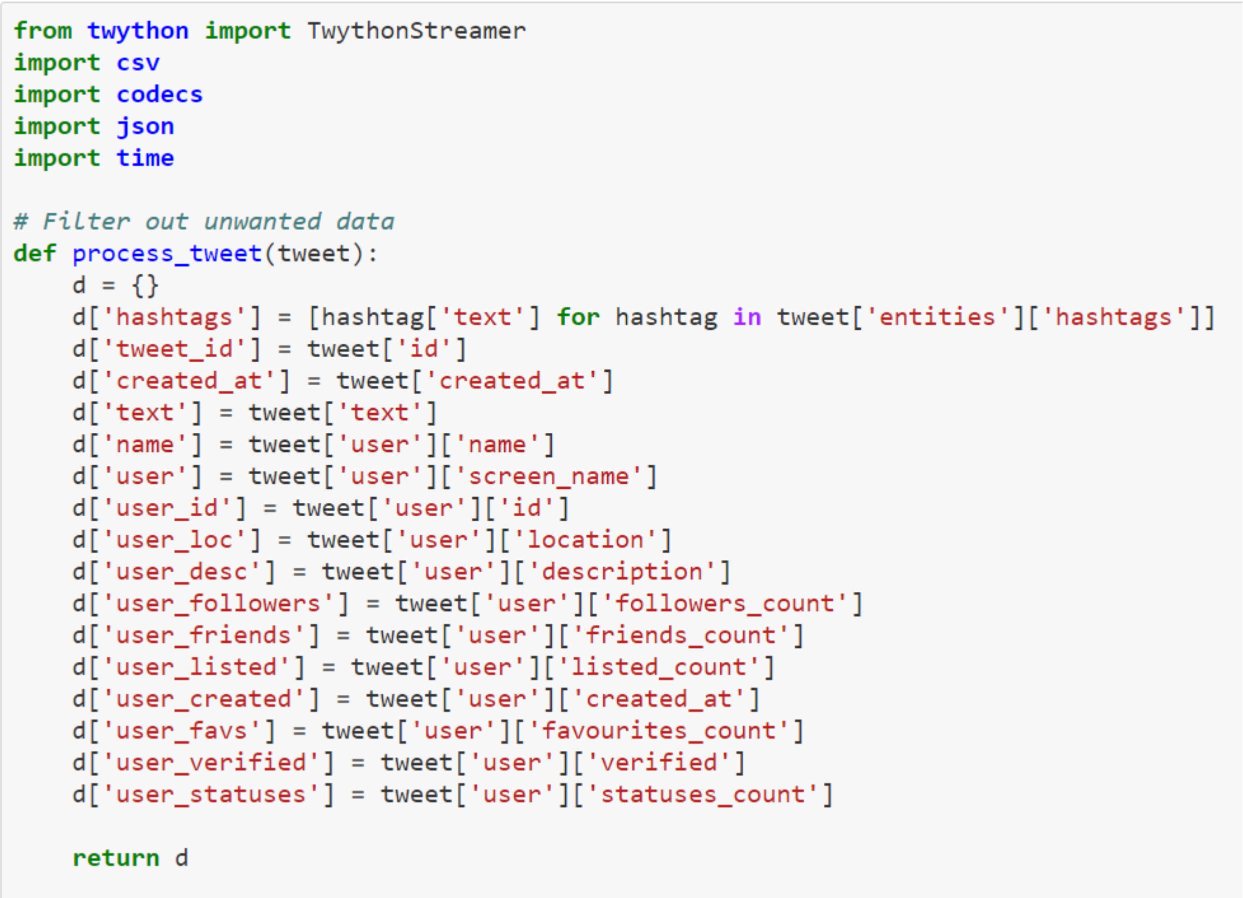
Team	Abbreviation
Anaheim Ducks	ANA
Arizona Coyotes	ARI
Boston Bruins	BOS
Buffalo Sabres	BUF
Calgary Flames	CGY
Carolina Hurricanes	CAR
Chicago Blackhawks	CHI
Colorado Avalanche	COL
Columbus Blue Jackets	CBJ
Dallas Stars	DAL
Detroit Red Wings	DET
Edmonton Oilers	EDM
Florida Panthers	FLA
Los Angeles Kings	LA
Minnesota Wild	MIN
Montreal Canadiens	MTL
Nashville Predators	NSH
New Jersey Devils	NJ
New York Islanders	NYI
New York Rangers	NYR
Ottawa Senators	OTT
Philadelphia Flyers	PHI
Pittsburgh Penguins	PIT
San Jose Sharks	SJ

St. Louis Blues	STL
Tampa Bay Lightning	TB
Toronto Maple Leafs	TOR
Vancouver Canucks	VAN
Vegas Golden Knights	VGK
Washington Capitals	WHS
Winnipeg Jets	WPG

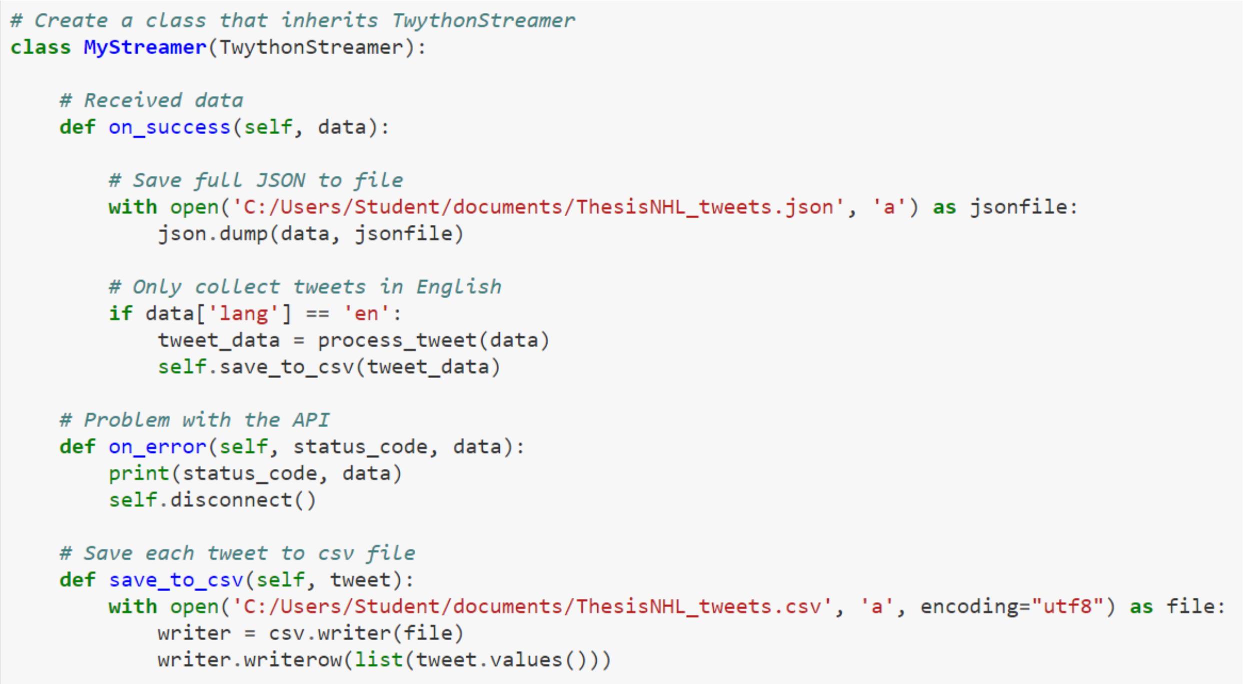
Appendix V: Tweet Collector



```
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
import json
import time
import os
import sys
from twython import Twython

twitter = Twython('0hRRJS1c3RgOuMUl92C8LIaBg',
'lBLr9WIaijeUinwCpQKo0fKstxAbz5HDoXdfGVTVF6BR42l7WI',
'1036474106-G5b2CZtsuCi0RgABfbPEI3uBhSPmXNwn40K78Jb',
'OThpz10mIpRwcJkhNTdlyZc62rWMCasPo11L09O35TbLl')

#twitter.get_home_timeline()
twitter.search(q='@NHL+OR+@NHLNetwork+OR+@TSNHockey')
```

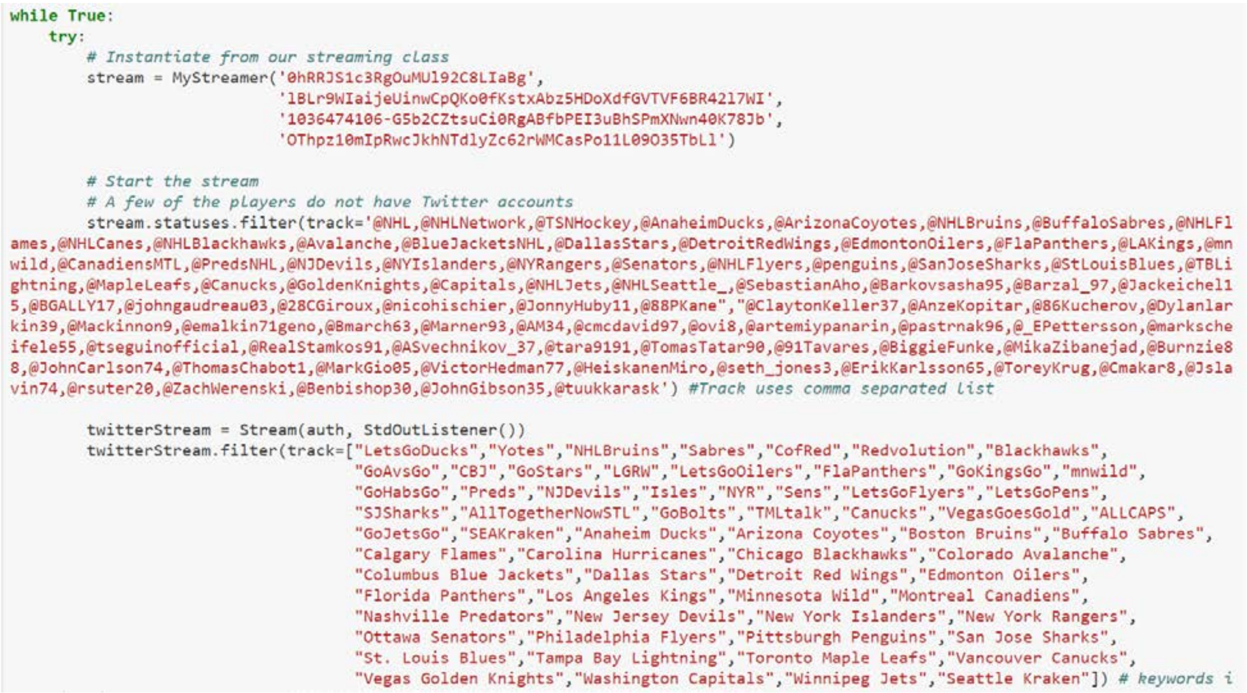


```
from twython import TwythonStreamer
import csv
import codecs
import json
import time

# Filter out unwanted data
def process_tweet(tweet):
    d = {}
    d['hashtags'] = [hashtag['text'] for hashtag in tweet['entities']['hashtags']]
    d['tweet_id'] = tweet['id']
    d['created_at'] = tweet['created_at']
    d['text'] = tweet['text']
    d['name'] = tweet['user']['name']
    d['user'] = tweet['user']['screen_name']
    d['user_id'] = tweet['user']['id']
    d['user_loc'] = tweet['user']['location']
    d['user_desc'] = tweet['user']['description']
    d['user_followers'] = tweet['user']['followers_count']
    d['user_friends'] = tweet['user']['friends_count']
    d['user_listed'] = tweet['user']['listed_count']
    d['user_created'] = tweet['user']['created_at']
    d['user_favs'] = tweet['user']['favourites_count']
    d['user_verified'] = tweet['user']['verified']
    d['user_statuses'] = tweet['user']['statuses_count']

    return d
```



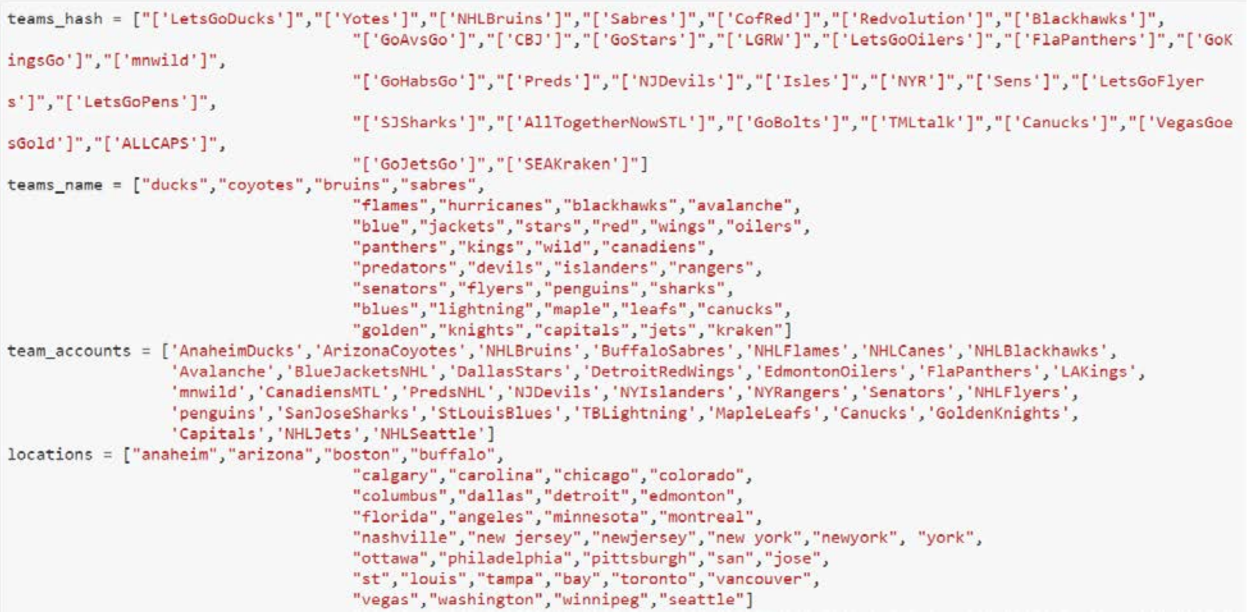
```
# Create a class that inherits TwythonStreamer
class MyStreamer(TwythonStreamer):

    # Received data
    def on_success(self, data):

        # Save full JSON to file
        with open('C:/Users/Student/documents/ThesisNHL_tweets.json', 'a') as jsonfile:
            json.dump(data, jsonfile)

        # Only collect tweets in English
        if data['lang'] == 'en':
            tweet_data = process_tweet(data)
            self.save_to_csv(tweet_data)

    # Problem with the API
    def on_error(self, status_code, data):
        print(status_code, data)
        self.disconnect()

    # Save each tweet to csv file
    def save_to_csv(self, tweet):
        with open('C:/Users/Student/documents/ThesisNHL_tweets.csv', 'a', encoding="utf8") as file:
            writer = csv.writer(file)
            writer.writerow(list(tweet.values()))
```

Appendix VI: Tweet Collector Criteria

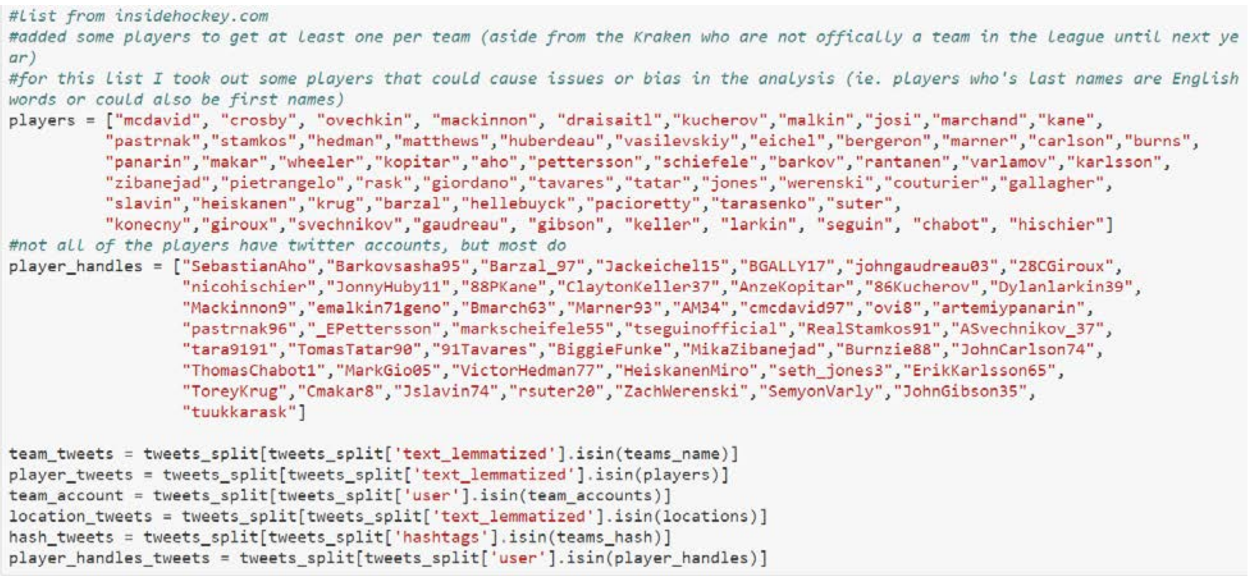


```
while True:
    try:
        # Instantiate from our streaming class
        stream = MyStreamer('0hRRJS1c3RgOuMUl92C8LIaBg',
                            '1BLr9WIaijeUinwCpQKo0fKstxAbz5HDoXdfGVTVF6BR42l7WI',
                            '1036474106-G5b2CZtsuCi0RgABfbPEI3uBhSPmXNwn40K78Jb',
                            'OThpz10mIpRwcJkhNTdlyZc62rWMCasPo11L09O35TbLl')

        # Start the stream
        # A few of the players do not have Twitter accounts
        stream.statuses.filter(track='@NHL,@NHLNetwork,@TSNHockey,@AnaheimDucks,@ArizonaCoyotes,@NHLBruins,@BuffaloSabres,@NHLFl
ames,@NHLCanes,@NHLBlackhawks,@Avalanche,@BlueJacketsNHL,@DallasStars,@DetroitRedWings,@EdmontonOilers,@FlaPanthers,@LAKings,@mn
wild,@CanadiensMTL,@PredsNHL,@NJDevils,@NYIslanders,@NYRangers,@Senators,@NHLFlyers,@penguins,@SanJoseSharks,@StLouisBlues,@TBLi
ghtning,@MapleLeafs,@Canucks,@GoldenKnights,@Capitals,@NHLJets,@NHLSeattle_,@SebastianAho,@Barkovsasha95,@Barzal_97,@Jackeichel1
5,@BGALLY17,@johngaudreau03,@28CGiroux,@nicohischier,@JonnyHuby11,@88PKane","@ClaytonKeller37,@AnzeKopitar,@86Kucherov,@Dylanlar
kin39,@Mackinnon9,@emalkin71geno,@Bmarch63,@Marner93,@AM34,@cmcdavid97,@ovi8,@artemiypanarin,@pastrnak96,@_EPettersson,@marksche
ifele55,@tseguinofficial,@RealStamkos91,@ASvechnikov_37,@tara9191,@TomasTatar90,@91Tavares,@BiggieFunke,@MikaZibanejad,@Burnzie8
8,@JohnCarlson74,@ThomasChabot1,@MarkGio05,@VictorHedman77,@HeiskanenMiro,@seth_jones3,@ErikKarlsson65,@ToreyKrug,@Cmakar8,@Jsla
vin74,@rsuter20,@ZachWerenski,@Benbishop30,@JohnGibson35,@tuukkarask') #Track uses comma separated list

        twitterStream = Stream(auth, StdOutListener())
        twitterStream.filter(track=["LetsGoDucks","Yotes","NHLBruins","Sabres","CofRed","Redvolution","Blackhawks",
                                    "GoAvsGo","CBJ","GoStars","LGRW","LetsGoOilers","FlaPanthers","GoKingsGo","mnwild",
                                    "GoHabsGo","Preds","NJDevils","Isles","NYR","Sens","LetsGoFlyers","LetsGoPens",
                                    "SJSharks","AllTogetherNowSTL","GoBolts","TMLtalk","Canucks","VegasGoesGold","ALLCAPS",
                                    "GoJetsGo","SEAKraken","Anaheim Ducks","Arizona Coyotes","Boston Bruins","Buffalo Sabres",
                                    "Calgary Flames","Carolina Hurricanes","Chicago Blackhawks","Colorado Avalanche",
                                    "Columbus Blue Jackets","Dallas Stars","Detroit Red Wings","Edmonton Oilers",
                                    "Florida Panthers","Los Angeles Kings","Minnesota Wild","Montreal Canadiens",
                                    "Nashville Predators","New Jersey Devils","New York Islanders","New York Rangers",
                                    "Ottawa Senators","Philadelphia Flyers","Pittsburgh Penguins","San Jose Sharks",
                                    "St. Louis Blues","Tampa Bay Lightning","Toronto Maple Leafs","Vancouver Canucks",
                                    "Vegas Golden Knights","Washington Capitals","Winnipeg Jets","Seattle Kraken"]) # keywords i
```

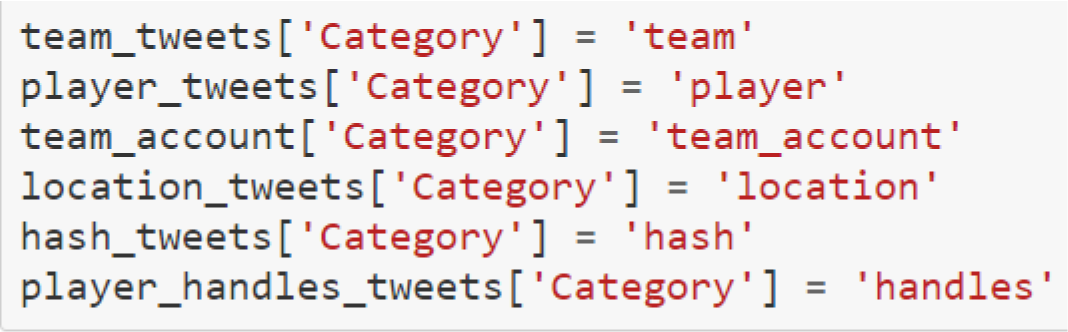
Appendix VII: Tweet Criteria Breakdown



```
teams_hash = ["['LetsGoDucks']","['Yotes']","['NHLBruins']","['Sabres']","['CofRed']","['Redvolution']","['Blackhawks']",
                                    "['GoAvsGo']","['CBJ']","['GoStars']","['LGRW']","['LetsGoOilers']","['FlaPanthers']","['GoK
ingsGo']","['mnwild']",
                                    "['GoHabsGo']","['Preds']","['NJDevils']","['Isles']","['NYR']","['Sens']","['LetsGoFlyer
s']","['LetsGoPens']",
                                    "['SJSharks']","['AllTogetherNowSTL']","['GoBolts']","['TMLtalk']","['Canucks']","['VegasGoe
sGold']","['ALLCAPS']",
                                    "['GoJetsGo']","['SEAKraken']"]
teams_name = ["ducks","coyotes","bruins","sabres",
                                    "flames","hurricanes","blackhawks","avalanche",
                                    "blue","jackets","stars","red","wings","oilers",
                                    "panthers","kings","wild","canadiens",
                                    "predators","devils","islanders","rangers",
                                    "senators","flyers","penguins","sharks",
                                    "blues","lightning","maple","leafs","canucks",
                                    "golden","knights","capitals","jets","kraken"]
team_accounts = ['AnaheimDucks','ArizonaCoyotes','NHLBruins','BuffaloSabres','NHLFlames','NHLCanes','NHLBlackhawks',
                 'Avalanche','BlueJacketsNHL','DallasStars','DetroitRedWings','EdmontonOilers','FlaPanthers','LAKings',
                 'mnwild','CanadiensMTL','PredsNHL','NJDevils','NYIslanders','NYRangers','Senators','NHLFlyers',
                 'penguins','SanJoseSharks','StLouisBlues','TBLightning','MapleLeafs','Canucks','GoldenKnights',
                 'Capitals','NHLJets','NHLSeattle']
locations = ["anaheim","arizona","boston","buffalo",
                                    "calgary","carolina","chicago","colorado",
                                    "columbus","dallas","detroit","edmonton",
                                    "florida","angeles","minnesota","montreal",
                                    "nashville","new jersey","newjersey","new york","newyork", "york",
                                    "ottawa","philadelphia","pittsburgh","san","jose",
                                    "st","louis","tampa","bay","toronto","vancouver",
                                    "vegas","washington","winnipeg","seattle"]
```

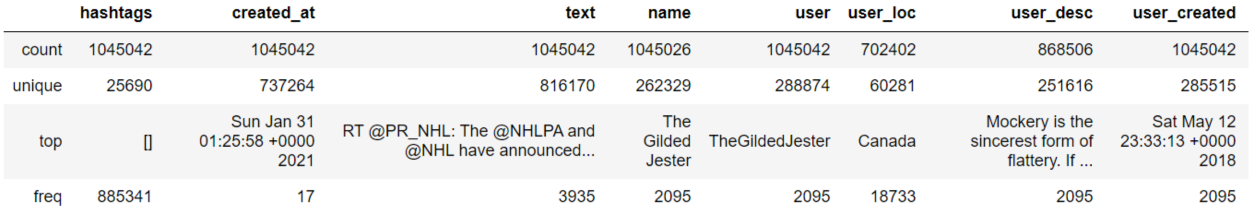


```
#list from insidehockey.com
#added some players to get at least one per team (aside from the Kraken who are not offically a team in the league until next ye
ar)
#for this list I took out some players that could cause issues or bias in the analysis (ie. players who's last names are English
words or could also be first names)
players = ["mcdavid", "crosby", "ovechkin", "mackinnon", "draisaitl","kucherov","malkin","josi","marchand","kane",
           "pastrnak","stamkos","hedman","matthews","huberdeau","vasilevskiy","eichel","bergeron","marner","carlson","burns",
           "panarin","makar","wheeler","kopitar","aho","pettersson","schiefele","barkov","rantanen","varlamov","karlsson",
           "zibanejad","pietrangelo","rask","giordano","tavares","tatar","jones","werenski","couturier","gallagher",
           "slavin","heiskanen","krug","barzal","hellebuyck","pacioretty","tarasenko","suter",
           "konecny","giroux","svechnikov","gaudreau", "gibson", "keller", "larkin", "seguin", "chabot", "hischier"]
#not all of the players have twitter accounts, but most do
player_handles = ["SebastianAho","Barkovsasha95","Barzal_97","Jackeichel15","BGALLY17","johngaudreau03","28CGiroux",
                  "nicohischier","JonnyHuby11","88PKane","ClaytonKeller37","AnzeKopitar","86Kucherov","Dylanlarkin39",
                  "Mackinnon9","emalkin71geno","Bmarch63","Marner93","AM34","cmcdavid97","ovi8","artemiypanarin",
                  "pastrnak96","_EPettersson","markscheifele55","tseguinofficial","RealStamkos91","ASvechnikov_37",
                  "tara9191","TomasTatar90","91Tavares","BiggieFunke","MikaZibanejad","Burnzie88","JohnCarlson74",
                  "ThomasChabot1","MarkGio05","VictorHedman77","HeiskanenMiro","seth_jones3","ErikKarlsson65",
                  "ToreyKrug","Cmakar8","Jslavin74","rsuter20","ZachWerenski","SemyonVarly","JohnGibson35",
                  "tuukkarask"]

team_tweets = tweets_split[tweets_split['text_lemmatized'].isin(teams_name)]
player_tweets = tweets_split[tweets_split['text_lemmatized'].isin(players)]
team_account = tweets_split[tweets_split['user'].isin(team_accounts)]
location_tweets = tweets_split[tweets_split['text_lemmatized'].isin(locations)]
hash_tweets = tweets_split[tweets_split['hashtags'].isin(teams_hash)]
player_handles_tweets = tweets_split[tweets_split['user'].isin(player_handles)]
```



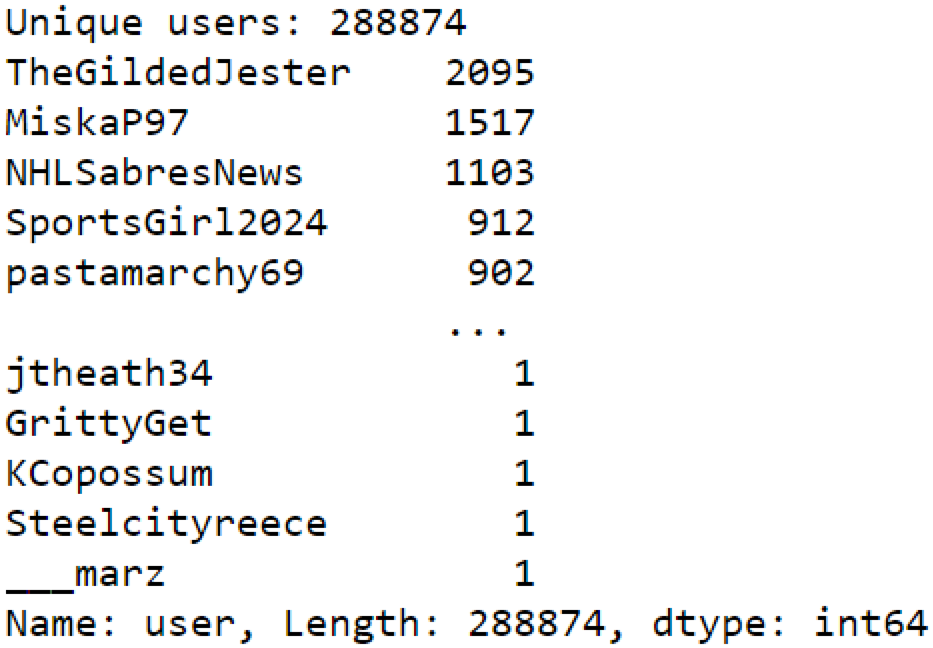
```
team_tweets['Category'] = 'team'
player_tweets['Category'] = 'player'
team_account['Category'] = 'team_account'
location_tweets['Category'] = 'location'
hash_tweets['Category'] = 'hash'
player_handles_tweets['Category'] = 'handles'
```

Appendix VIII: Tweet Summary Statistics



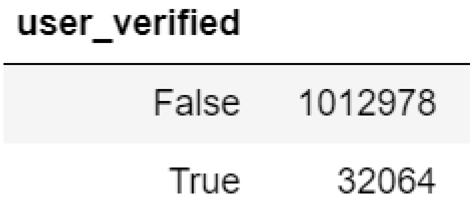
	hashtags	created_at	text	name	user	user_loc	user_desc	user_created
count	1045042	1045042	1045042	1045026	1045042	702402	868506	1045042
unique	25690	737264	816170	262329	288874	60281	251616	285515
top	[]	Sun Jan 31 01:25:58 +0000 2021	RT @PR_NHL: The @NHLPA and @NHL have announced...	The Gilded Jester	TheGildedJester	Canada	Mockery is the sincerest form of flattery. If ...	Sat May 12 23:33:13 +0000 2018
freq	885341	17	3935	2095	2095	18733	2095	2095

Appendix IX: Unique Users



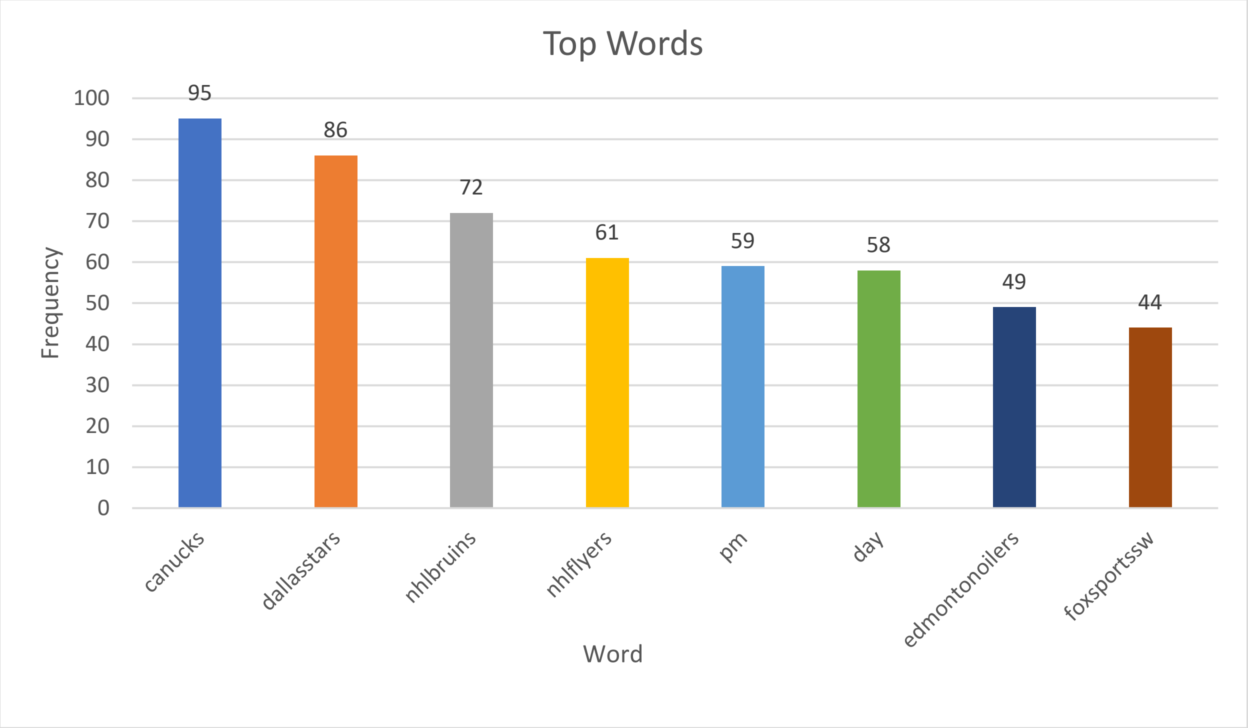
```
Unique users: 288874
TheGildedJester    2095
MiskaP97           1517
NHLSabresNews      1103
SportsGirl2024      912
pastamarchy69       902
                   ...
jtheath34             1
GrittyGet             1
KCopossum             1
Steelcityreece        1
___marz               1
Name: user, Length: 288874, dtype: int64
```

Appendix X: Verified Users

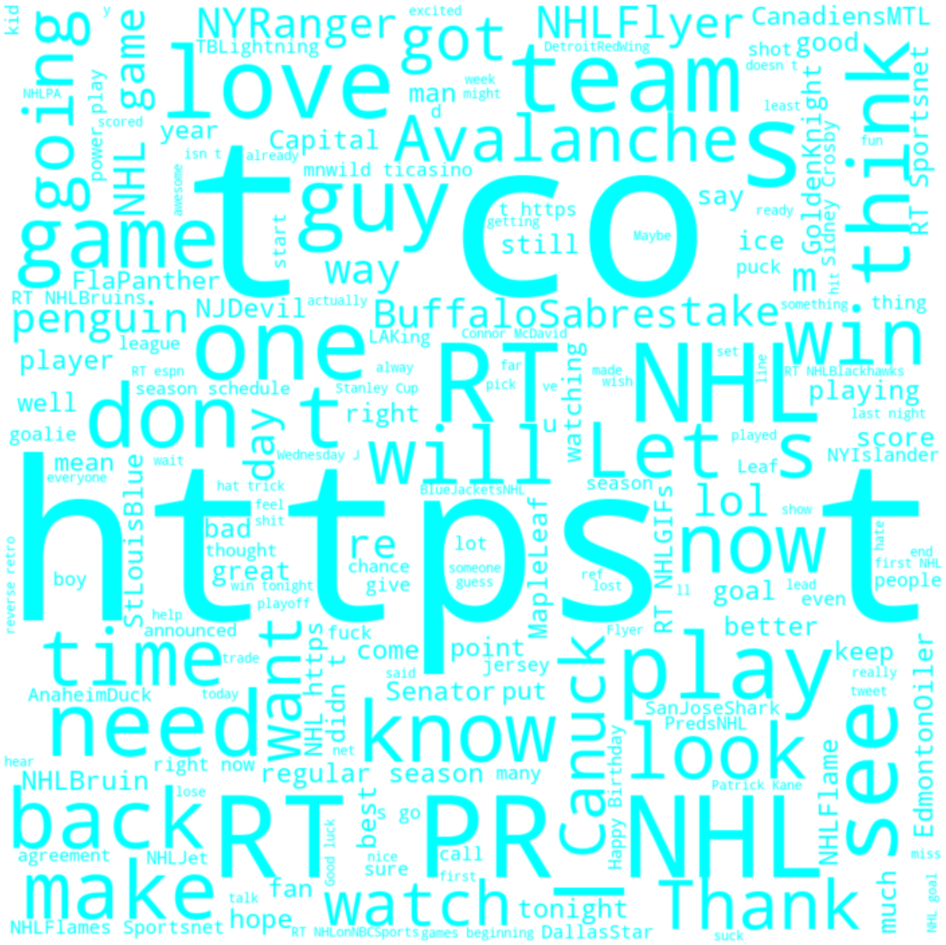


user_verified	
False	1012978
True	32064

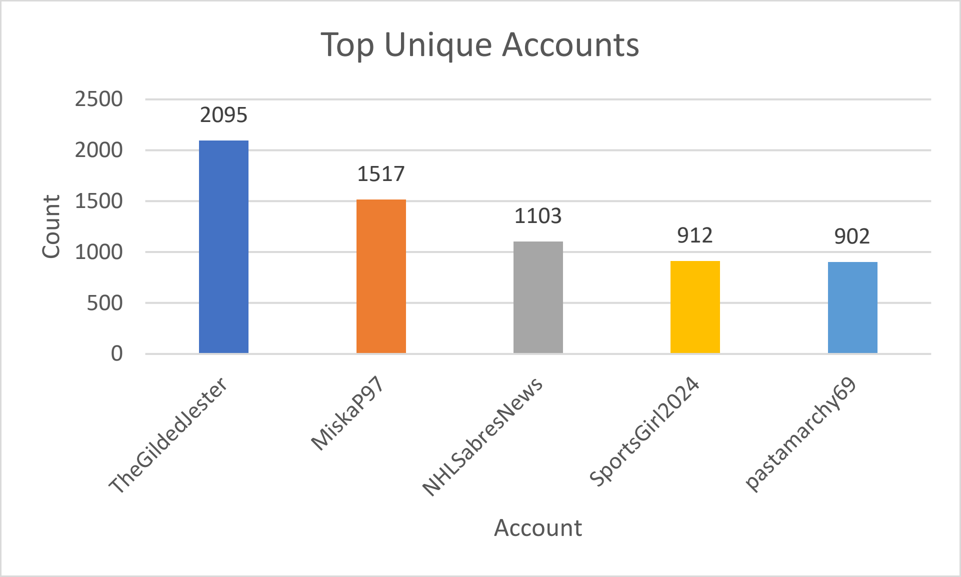
Appendix XI: Top Words



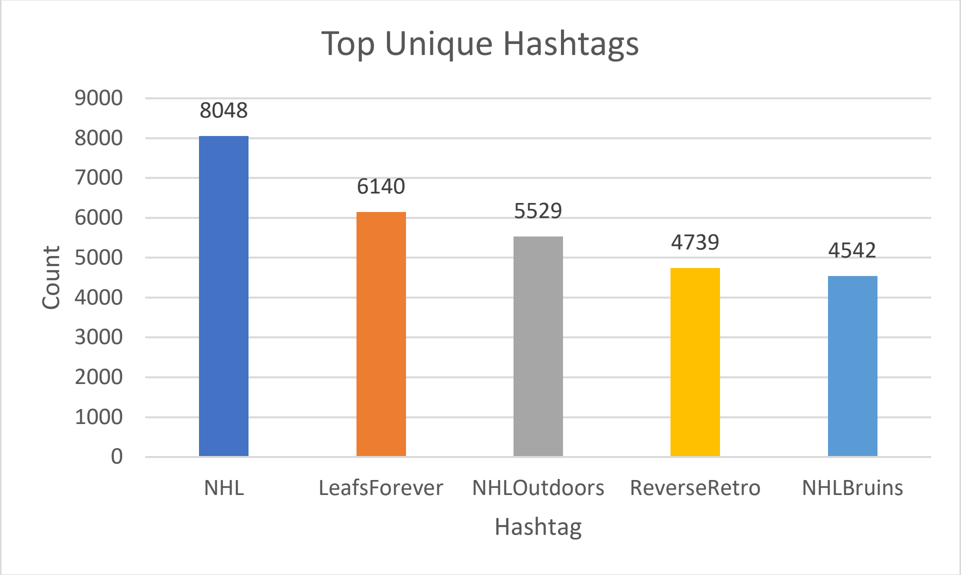
Appendix XII: Top Words Cloud



Appendix XIII: Top Accounts



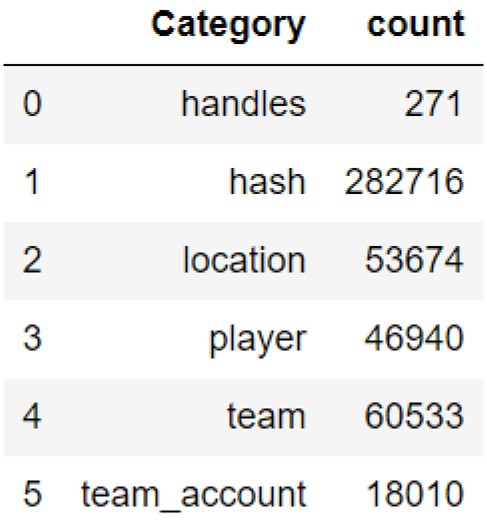
Appendix XIV: Top Hashtags



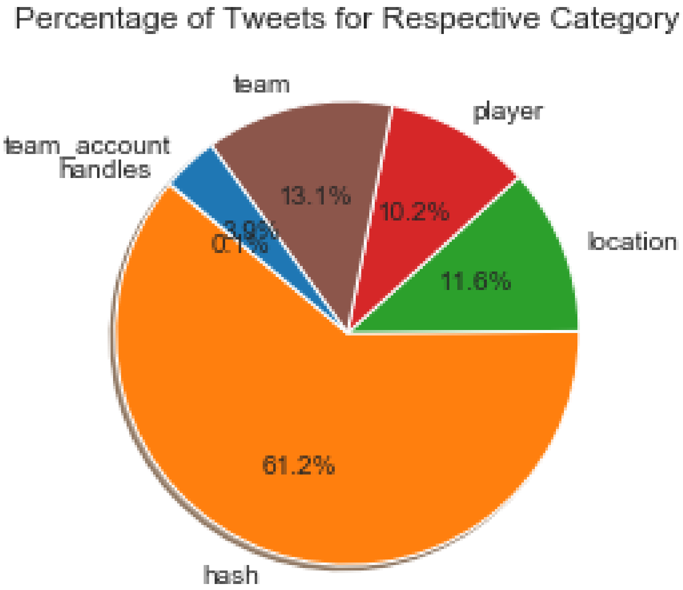
Appendix XV: Popular Accounts

Account	Count	Verified
TheGildedJester	2095	0
MiskaP97	1517	0
NHLSabresNews	1103	0
SportsGirl2024	912	0
pastamarchy69	902	0

Appendix XVI: Tweet Category Breakdowns



	Category	count
0	handles	271
1	hash	282716
2	location	53674
3	player	46940
4	team	60533
5	team_account	18010



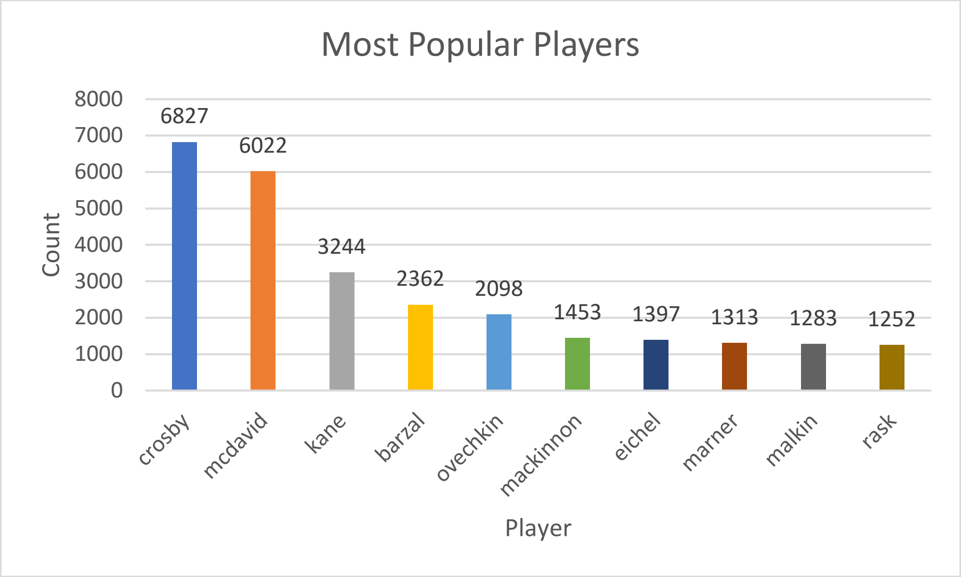
Appendix XVII: Popular Players

crosby	6827
mcdavid	6022
kane	3244
barzal	2362
ovechkin	2098
mackinnon	1453
eichel	1397
marner	1313
malkin	1283
rask	1252
pastrnak	1143
suter	1063
marchand	1007
draisaitl	999
panarin	977
bergeron	801
giroux	781

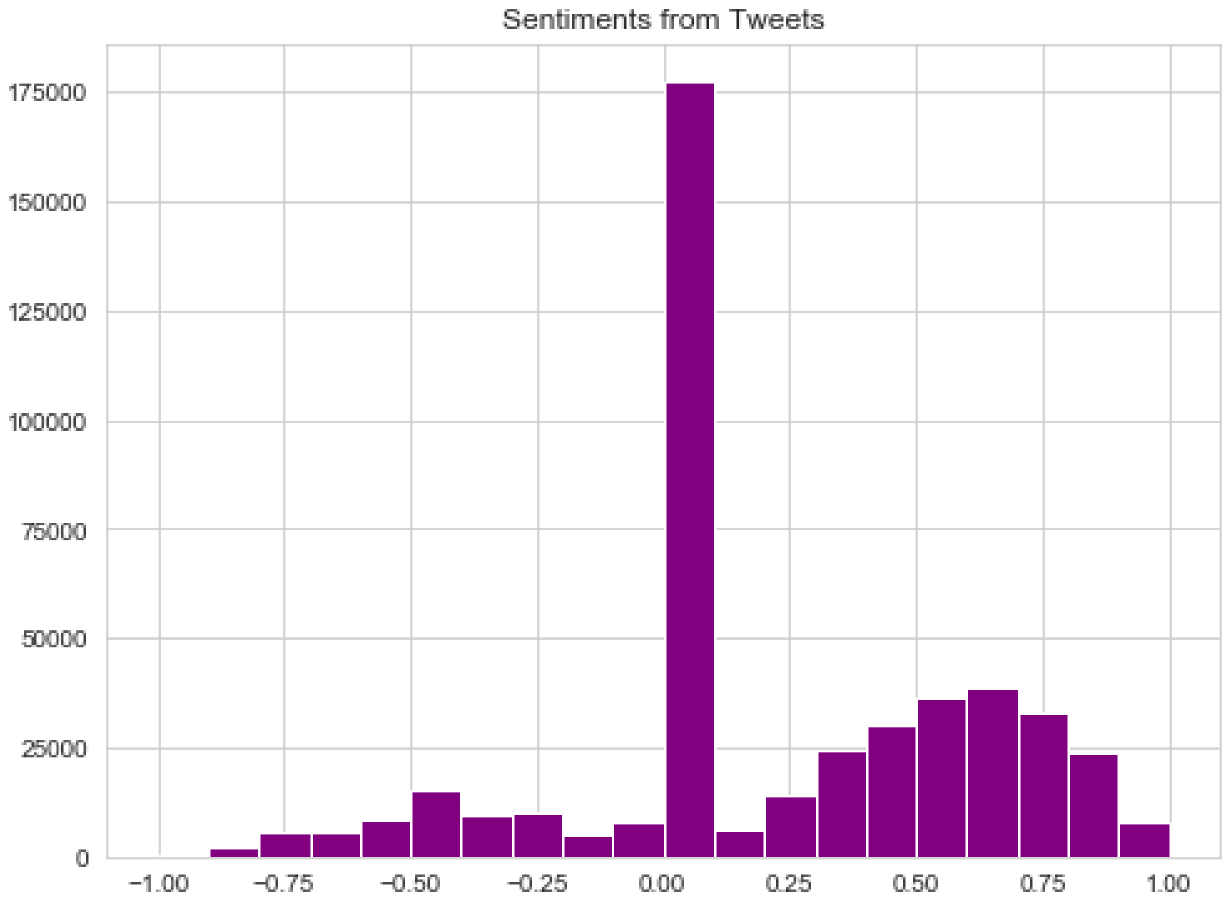
jones	702
larkin	695
makar	639
pettersson	622
tavares	572
wheeler	553
gaudreau	480
huberdeau	474
zibanejad	470
stamkos	467
gibson	458
hedman	454
vasilevskiy	415
barkov	410
pacioretty	387
konecny	380
rantanen	341
tarasenko	328
varlamov	326
kopitar	326
karlsson	295
gallagher	273
josi	267
krug	260
hellebuyck	254
couturier	243
chabot	234
pietrangelo	204
tatar	152
heiskanen	134
carlson	133

giordano	128
keller	128
kucherov	122
aho	114
hischier	109
svechnikov	104
werenski	91
seguin	88
slavin	76
schiefele	10

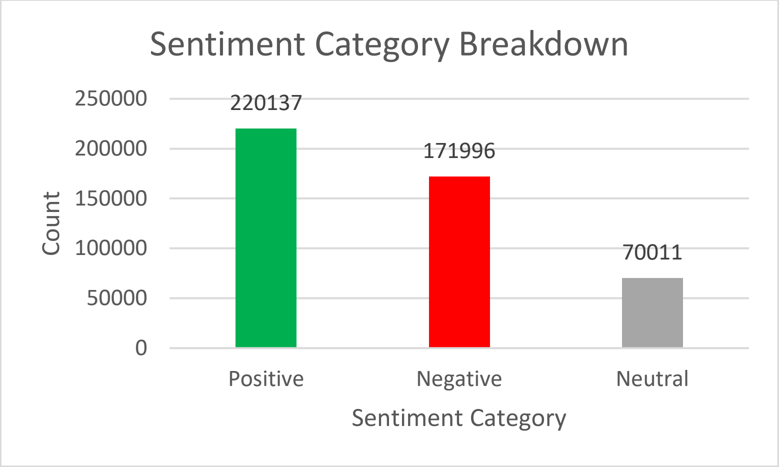
Appendix XVIII: Most Popular Players

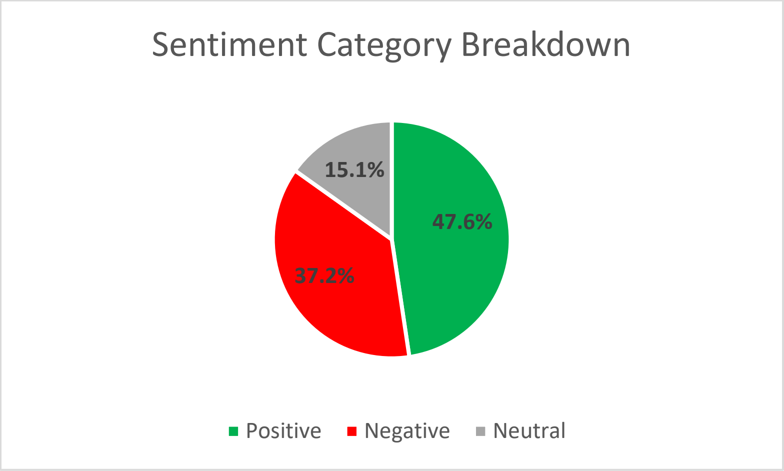


Appendix XIX: Sentiment Spread with Neutrals

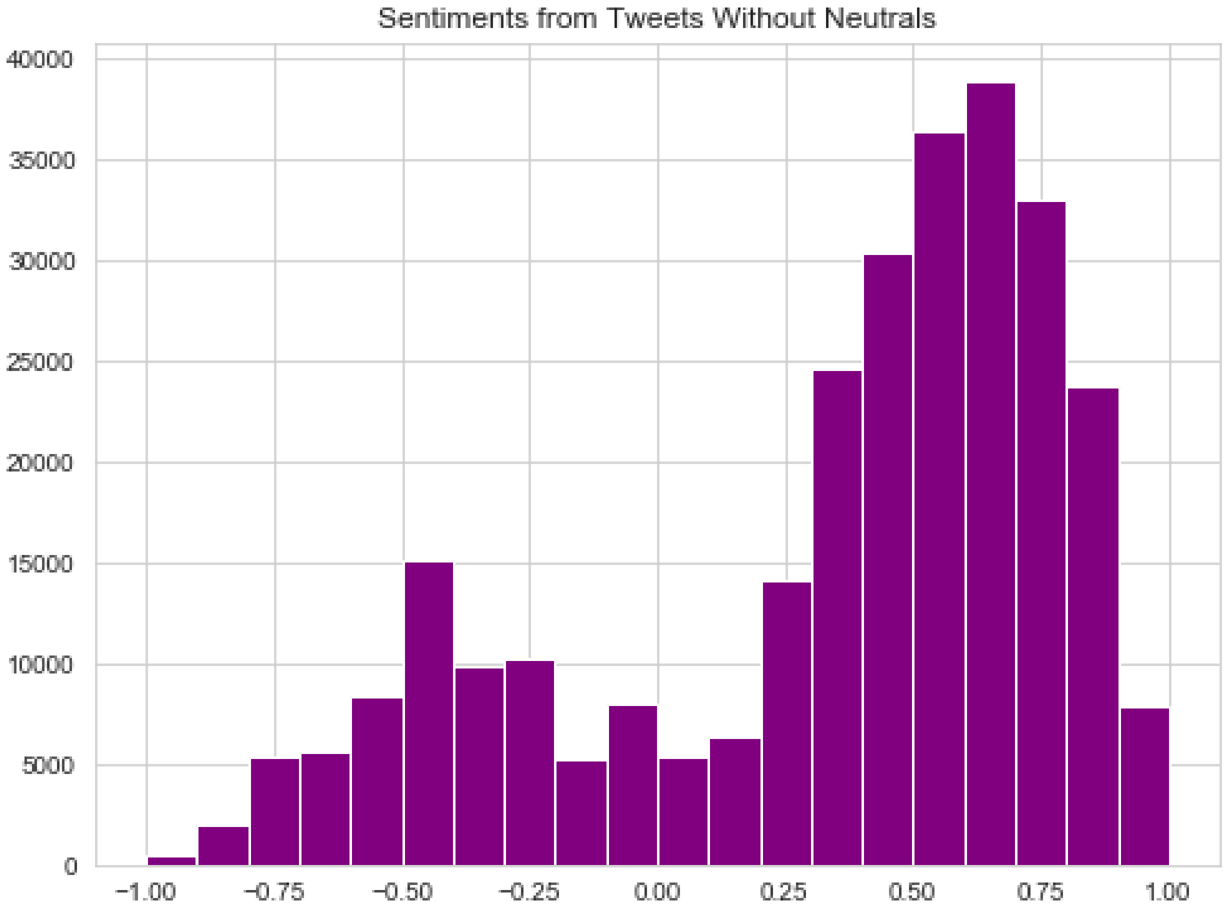


Appendix XX: Sentiment Category Breakdown

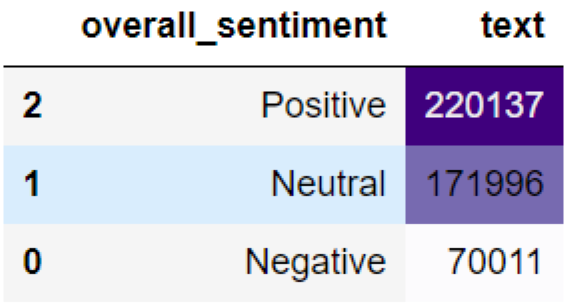




Appendix XXI: Sentiment of Tweets Spread without Neutrals

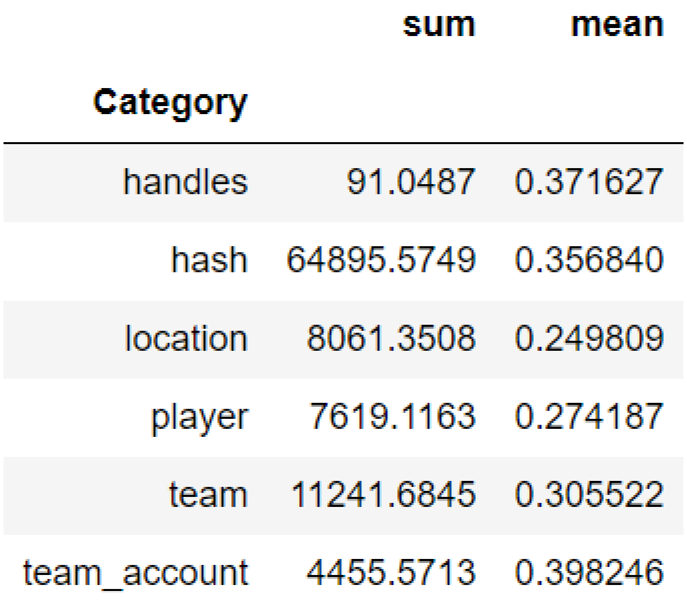


Appendix XXII: Overall Sentiment Breakdown



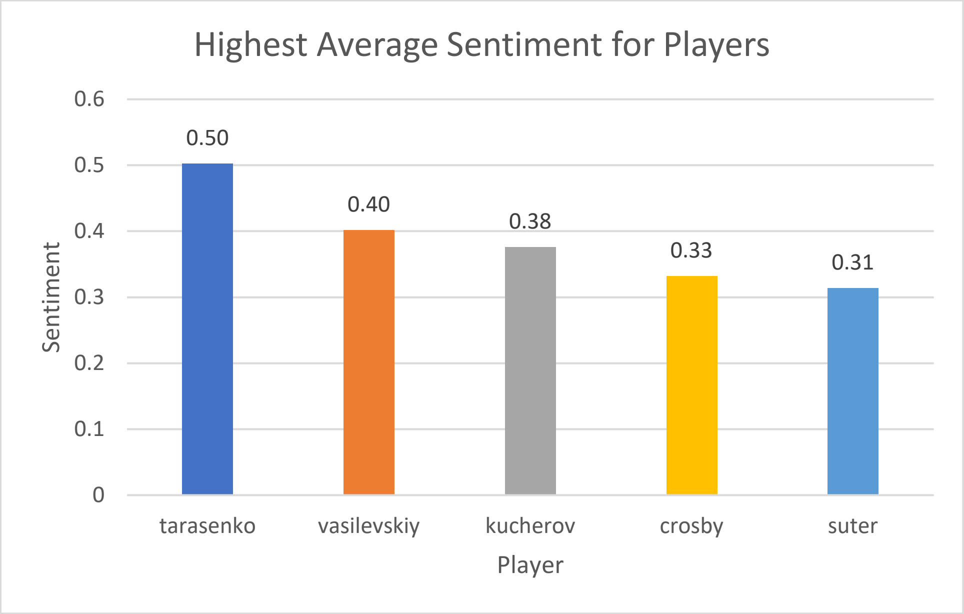
	overall_sentiment	text
2	Positive	220137
1	Neutral	171996
0	Negative	70011

Appendix XXIII: Tweet Category Statistics

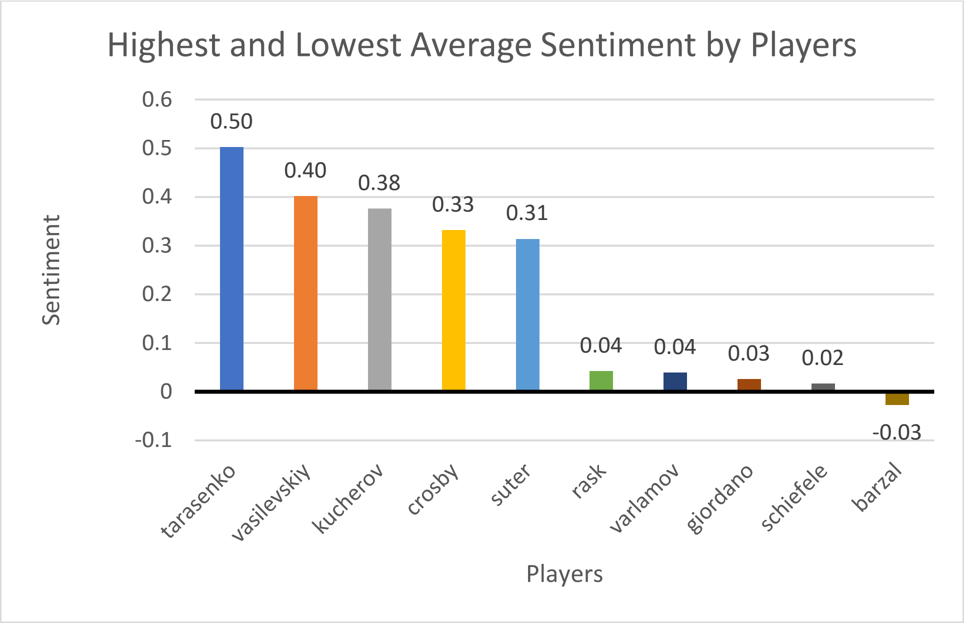


	sum	mean
Category		
handles	91.0487	0.371627
hash	64895.5749	0.356840
location	8061.3508	0.249809
player	7619.1163	0.274187
team	11241.6845	0.305522
team_account	4455.5713	0.398246

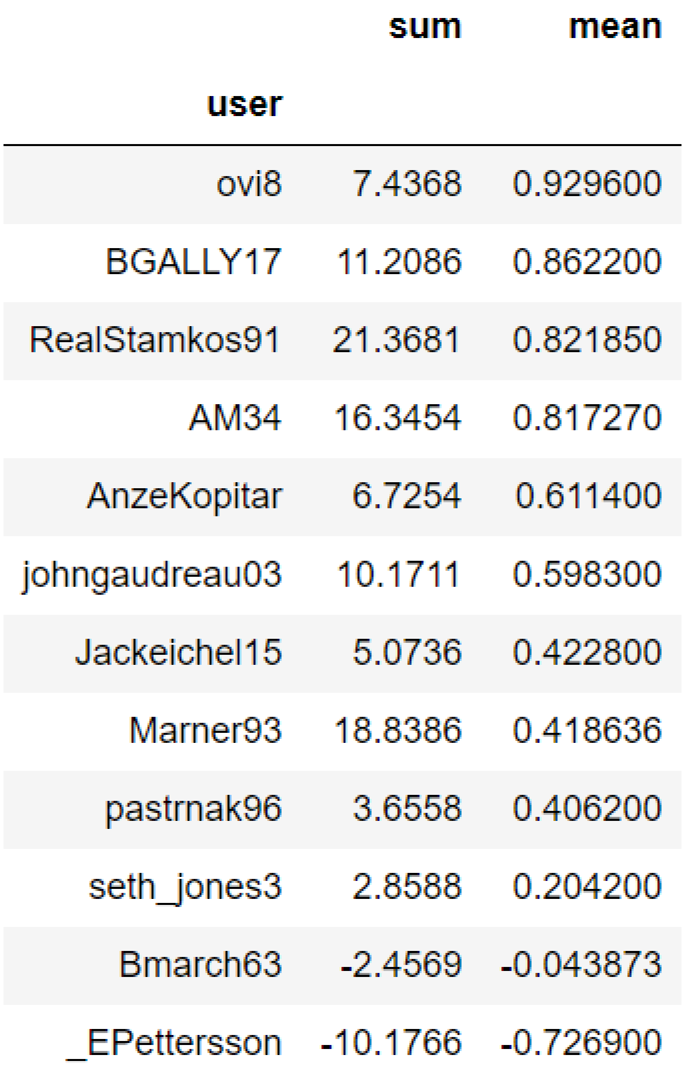
Appendix XXIV: Highest Average Sentiment for Players



Appendix XXV: Highest and Lowest Average Sentiment for Players

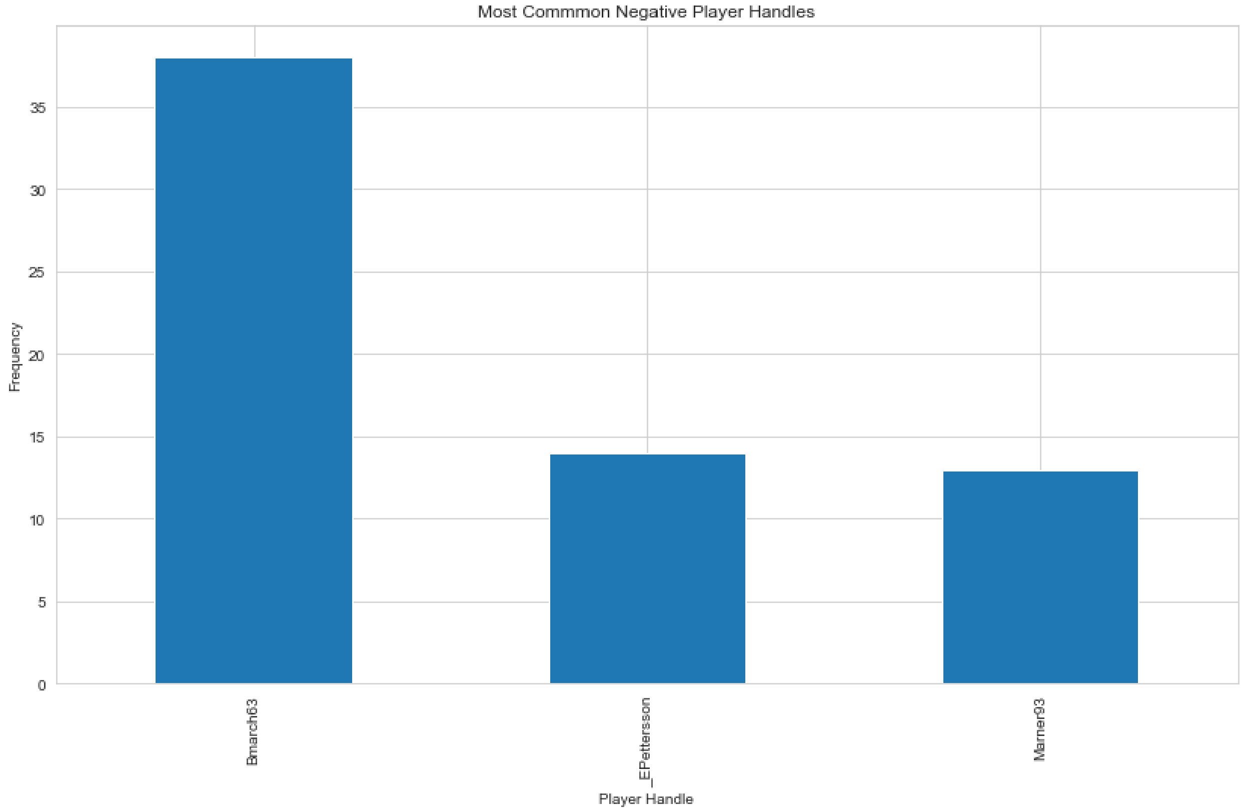


Appendix XXVI: Highest Sentiment from Player Accounts

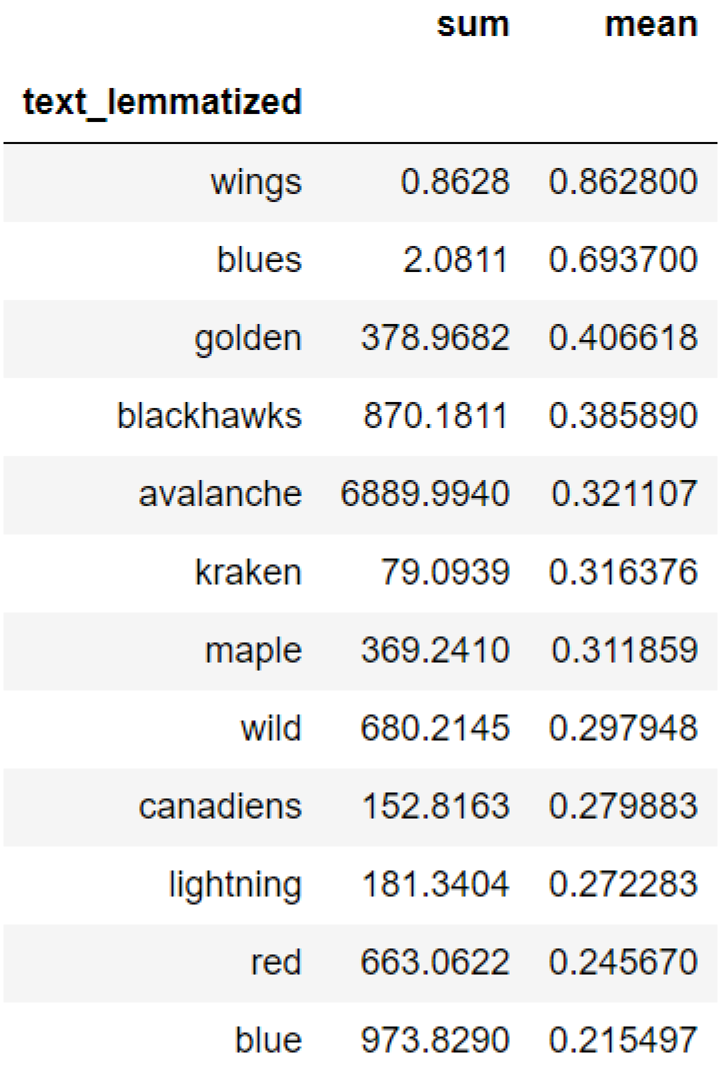


	sum	mean
user		
ovi8	7.4368	0.929600
BGALLY17	11.2086	0.862200
RealStamkos91	21.3681	0.821850
AM34	16.3454	0.817270
AnzeKopitar	6.7254	0.611400
johngaudreau03	10.1711	0.598300
Jackeichel15	5.0736	0.422800
Marner93	18.8386	0.418636
pastrnak96	3.6558	0.406200
seth_jones3	2.8588	0.204200
Bmarch63	-2.4569	-0.043873
_EPettersson	-10.1766	-0.726900

Appendix XXVII: Most Common Negative Player Accounts

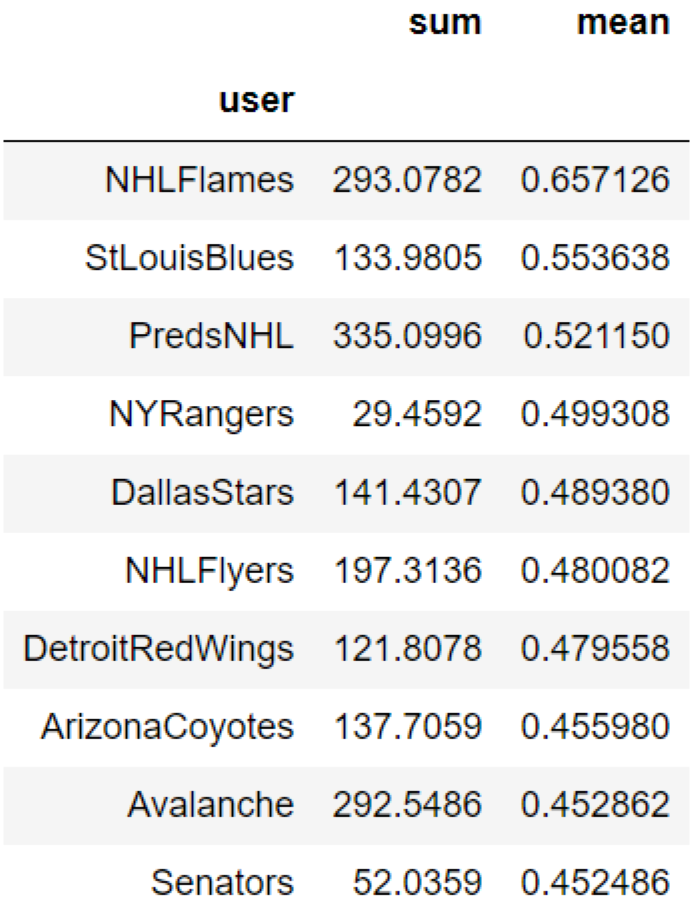


Appendix XXVIII: Most Positive Team Names



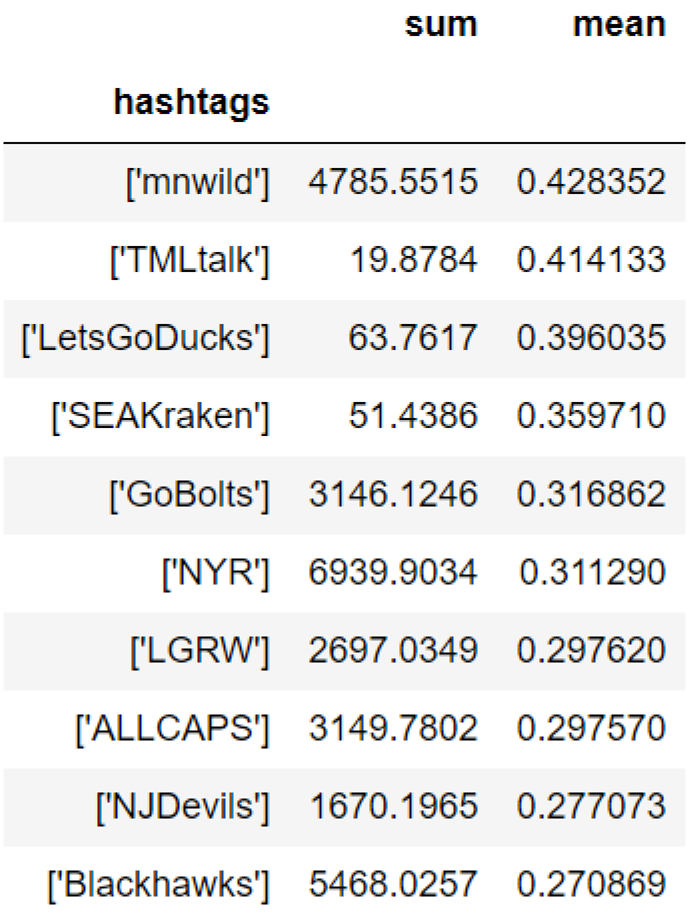
	sum	mean
text_lemmatized		
wings	0.8628	0.862800
blues	2.0811	0.693700
golden	378.9682	0.406618
blackhawks	870.1811	0.385890
avalanche	6889.9940	0.321107
kraken	79.0939	0.316376
maple	369.2410	0.311859
wild	680.2145	0.297948
canadiens	152.8163	0.279883
lightning	181.3404	0.272283
red	663.0622	0.245670
blue	973.8290	0.215497

Appendix XXIX: Most Positive Team Accounts



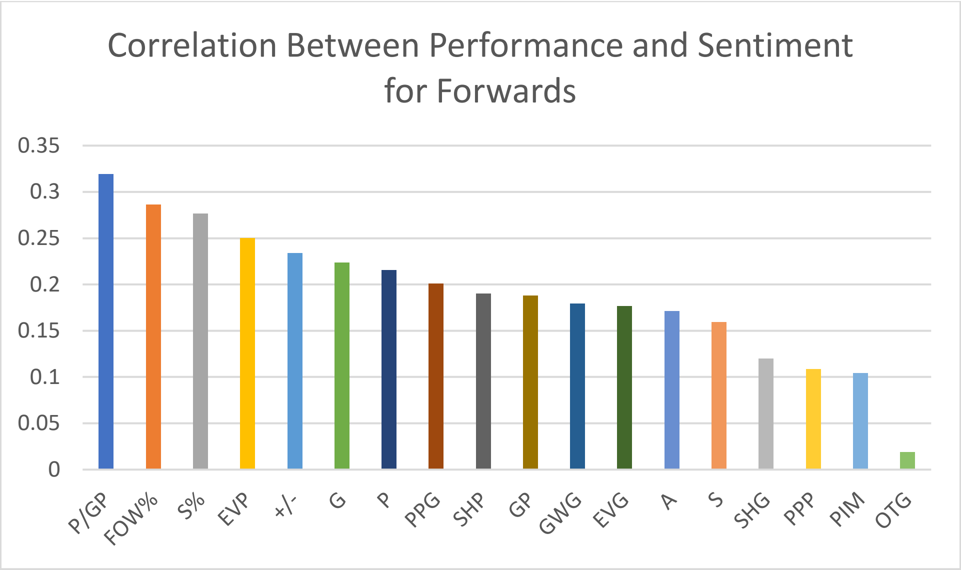
	sum	mean
user		
NHLFlames	293.0782	0.657126
StLouisBlues	133.9805	0.553638
PredsNHL	335.0996	0.521150
NYRangers	29.4592	0.499308
DallasStars	141.4307	0.489380
NHLFlyers	197.3136	0.480082
DetroitRedWings	121.8078	0.479558
ArizonaCoyotes	137.7059	0.455980
Avalanche	292.5486	0.452862
Senators	52.0359	0.452486

Appendix XXX: Most Positive Team Hashtags



hashtags	sum	mean
['mnwild']	4785.5515	0.428352
['TMLtalk']	19.8784	0.414133
['LetsGoDucks']	63.7617	0.396035
['SEAKraken']	51.4386	0.359710
['GoBolts']	3146.1246	0.316862
['NYR']	6939.9034	0.311290
['LGRW']	2697.0349	0.297620
['ALLCAPS']	3149.7802	0.297570
['NJDevils']	1670.1965	0.277073
['Blackhawks']	5468.0257	0.270869

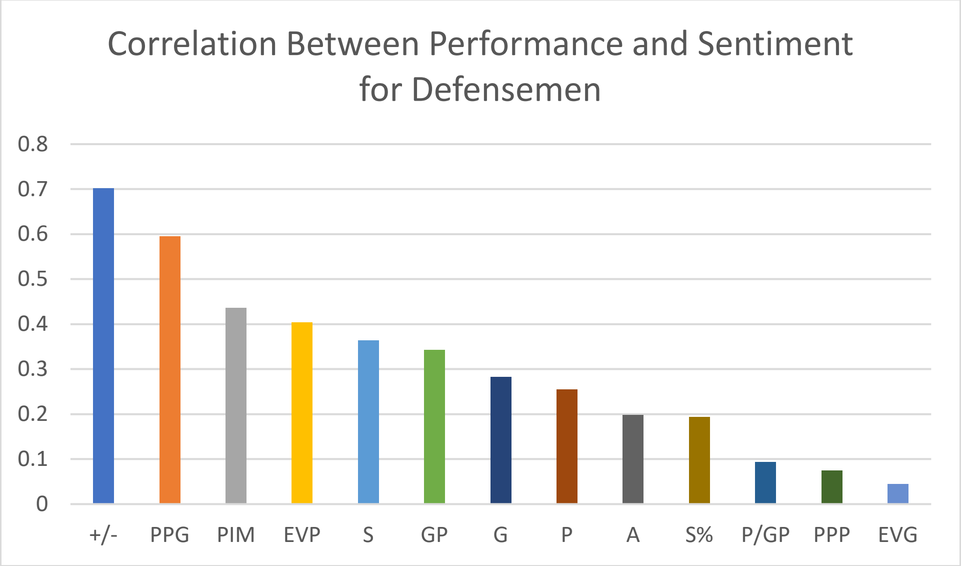
Appendix XXXI: Correlation Between Performance and Sentiment for Forwards



	W1	W2	W3	W4	W5	W6	W7	W8	All
+/-	0.047214	-0.14546	-0.00438	-0.04394	-0.06746	0.210011	0.071948	-0.0265	-0.233922
A	0.073004	-0.10964	0.018021	-0.01755	-0.07425	0.170193	-0.20087	-0.0947	-0.171233
EVG	-0.09542	-0.37667	0.385787	-0.08996	-0.26805	0.135295	0.27928	-0.13516	-0.176547
EVP	0.056154	-0.2343	0.184843	-0.10246	-0.22412	0.150826	0.03349	-0.1984	-0.249911
FOW%	0.285104	-0.46481	0	-0.26726	-0.23617	0.094341	0.14267	0.122122	-0.286315
G	-0.15038	-0.3623	0.362155	-0.21546	-0.26768	0.156897	0.13961	-0.13524	-0.223435
GP	-0.0952	-0.17813	0.403497	-0.12064	-0.12604	0.122263	0.027596	-0.05919	-0.188011
GWG	0.038409	0.046105	0.116147	-0.16149	-0.03311	0.280389	0.163715	0.06767	-0.179537
mean	1	1	1	1	1	1	1	1	1
OTG	0.062254	0	0.149945	-0.15951	-0.13673	0.038435	0.006019	-0.02241	0.018781
P	-0.03734	-0.26607	0.158015	-0.13535	-0.18249	0.193115	-0.08622	-0.14373	-0.21561
P/GP	-0.00463	-0.25333	0.207565	-0.27853	-0.16392	0.210149	0.064883	-0.15244	-0.31942
PIM	0.213132	0.150227	0.319118	0.143803	-0.00079	-0.05211	-0.00053	0.142026	0.103953
PPG	-0.15449	-0.15194	0.187549	-0.29202	-0.13438	0.132818	-0.22647	-0.11436	-0.200879
PPP	-0.13548	-0.23158	0.090049	-0.12328	0.008241	0.196684	-0.23349	-0.03639	-0.108358
S	0.089527	-0.27571	0.464134	-0.19474	-0.25596	0.144445	-0.00491	-0.13576	-0.159291
S%	-0.23613	-0.19431	0.38146	-0.12052	-0.03852	0.222951	0.237252	-0.08323	-0.276715
SHG	0.164399	-0.07335	0	0	0	0	0	0.227246	-0.12012

SHP	0.072844	-0.07335	0	0	0.00392	0	0.034239	0.227246	-0.189865

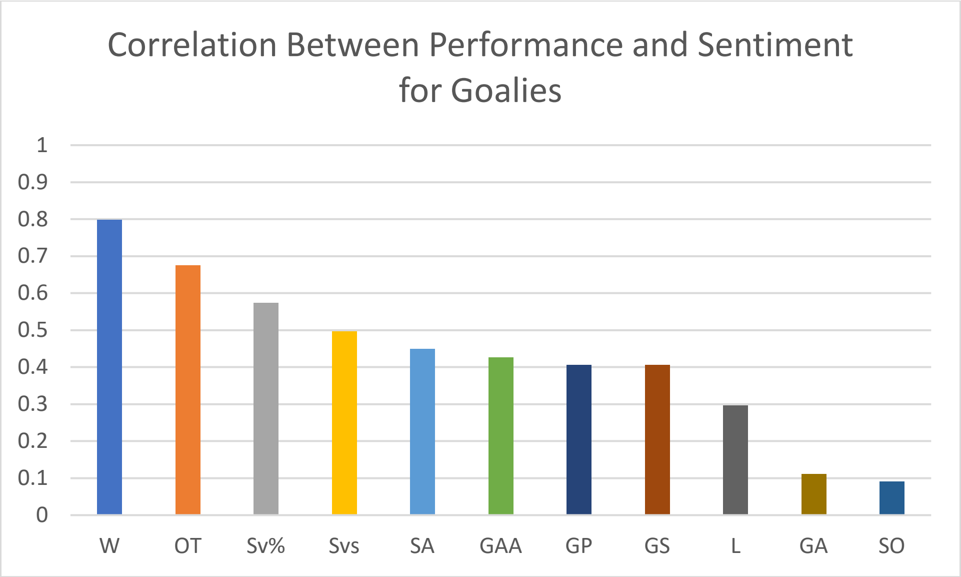
Appendix XXXII: Correlation Between Performance and Sentiment for Defensemen



	W1	**W2**	**W3**	**W4**	**W5**	**W6**	**W7**	**W8**	**All**
+/-	0.047214	-0.14546	-0.00438	-0.04394	-0.06746	0.210011	0.071948	-0.0265	-0.233922
A	0.073004	-0.10964	0.018021	-0.01755	-0.07425	0.170193	-0.20087	-0.0947	-0.171233
EVG	-0.09542	-0.37667	0.385787	-0.08996	-0.26805	0.135295	0.27928	-0.13516	-0.176547
EVP	0.056154	-0.2343	0.184843	-0.10246	-0.22412	0.150826	0.03349	-0.1984	-0.249911
FOW%	0.285104	-0.46481	0	-0.26726	-0.23617	0.094341	0.14267	0.122122	-0.286315
G	-0.15038	-0.3623	0.362155	-0.21546	-0.26768	0.156897	0.13961	-0.13524	-0.223435
GP	-0.0952	-0.17813	0.403497	-0.12064	-0.12604	0.122263	0.027596	-0.05919	-0.188011

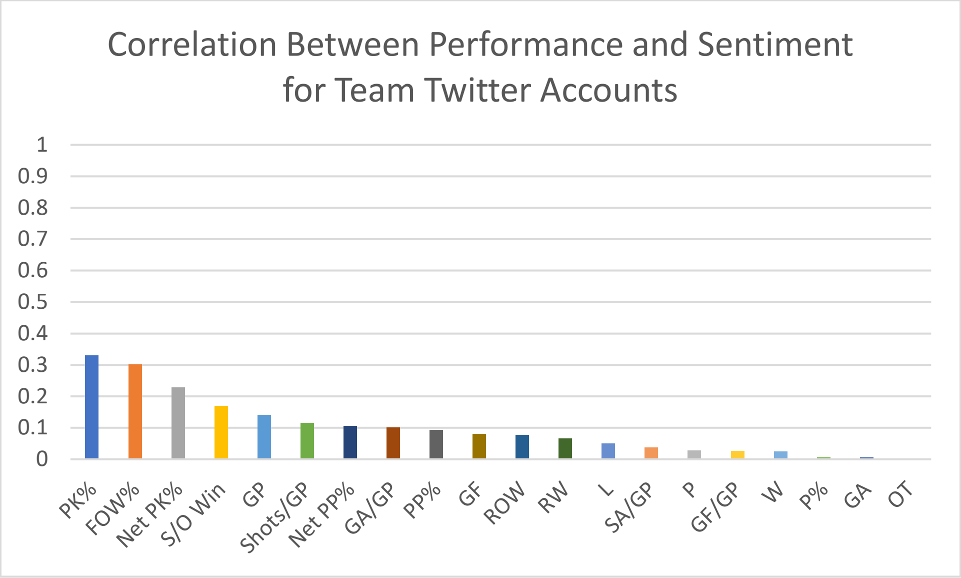
GWG	0.038409	0.046105	0.116147	-0.16149	-0.03311	0.280389	0.163715	0.06767	-0.179537
mean	1	1	1	1	1	1	1	1	1
OTG	0.062254	0	0.149945	-0.15951	-0.13673	0.038435	0.006019	-0.02241	0.018781
P	-0.03734	-0.26607	0.158015	-0.13535	-0.18249	0.193115	-0.08622	-0.14373	-0.21561
P/GP	-0.00463	-0.25333	0.207565	-0.27853	-0.16392	0.210149	0.064883	-0.15244	-0.31942
PIM	0.213132	0.150227	0.319118	0.143803	-0.00079	-0.05211	-0.00053	0.142026	0.103953
PPG	-0.15449	-0.15194	0.187549	-0.29202	-0.13438	0.132818	-0.22647	-0.11436	-0.200879
PPP	-0.13548	-0.23158	0.090049	-0.12328	0.008241	0.196684	-0.23349	-0.03639	-0.108358
S	0.089527	-0.27571	0.464134	-0.19474	-0.25596	0.144445	-0.00491	-0.13576	-0.159291
S%	-0.23613	-0.19431	0.38146	-0.12052	-0.03852	0.222951	0.237252	-0.08323	-0.276715
SHG	0.164399	-0.07335	0	0	0	0	0	0.227246	-0.12012
SHP	0.072844	-0.07335	0	0	0.00392	0	0.034239	0.227246	-0.189865

Appendix XXXIII: Correlation Between Performance and Sentiment for Goalies



	W1	W2	W3	W4	W5	W6	W7	W8	All
GA	0.658282	0.672902	-0.11098	0.538281	0.267681	0.083068	-0.65832	0.18525	-0.111908
GAA	0.769435	0.085918	-0.34589	0.292622	0.825091	-0.36135	-0.85426	0.049732	-0.426671
GP	0.101507	0.906133	-0.00204	-0.51025	-0.35989	0.301487	0.814784	0.39987	0.405854
GS	0.101507	0.906133	-0.00204	-0.51025	-0.35989	0.301487	0.814784	0.39987	0.405854
L	0.508676	-0.63496	0.129441	0.647844	0.190937	0.211373	-0.90982	-0.68999	-0.297291
mean	1	1	1	1	1	1	1	1	1
OT	0.101507	0.196379	0.040257	-0.07109	-0.19094	0	0.144954	-0.39698	-0.675787
SA	0.397565	0.51254	0.283629	-0.46907	0.040868	0.38218	0.718049	0.306687	0.450225
SO	-0.35864	0.196379	0	0.218609	-0.55756	0	0.658322	0	0.091688
Sv%	-0.52273	-0.24414	0.56825	-0.27275	-0.76133	0.570568	0.899606	-0.17367	0.57412
Svs	0.350934	0.419638	0.30578	-0.47409	0.023206	0.410564	0.830081	0.332016	0.496623
W	-0.50868	0.909305	-0.19997	-0.54267	-0.35989	0.025134	0.758837	0.714256	0.798594

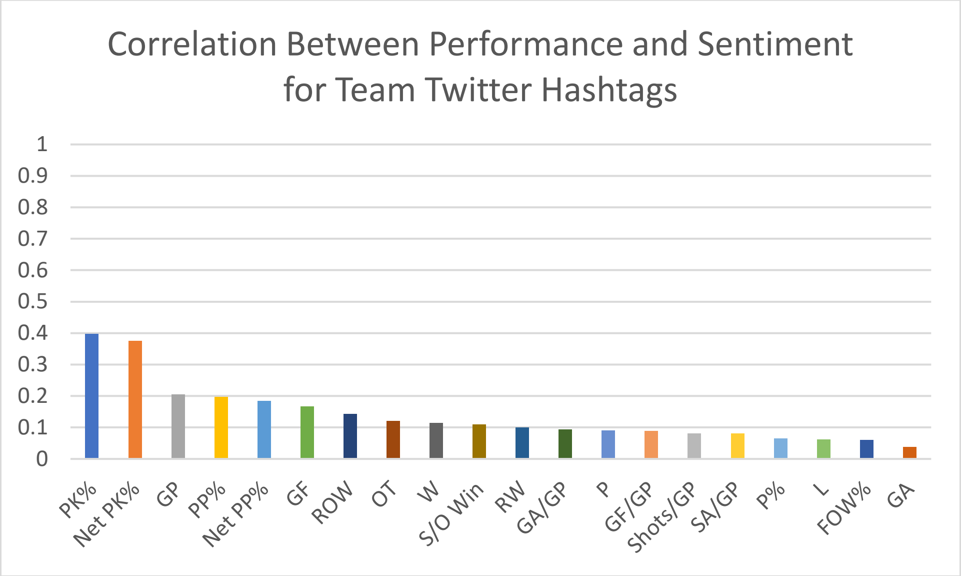
Appendix XXXIV: Correlation Between Performance and Sentiment for Accounts



	W1	W2	W3	W4	W5	W6	W7	W8	All
FOW%	-0.06778	0.01435	-0.24372	-0.09902	0.093332	0.090348	0.365985	0.229782	-0.30142
GA	0.085408	-0.14117	0.291861	0.223016	0.33815	0.023296	-0.2947	0.196544	0.005319
GA/GP	0.074773	-0.32114	0.064182	0.15584	0.492975	0.111228	-0.21943	0.021405	0.100666
GF	0.149593	0.20832	0.113753	-0.05204	-0.02921	-0.17502	-0.13703	0.204968	-0.080643
GF/GP	0.17318	0.148798	-0.00586	-0.09716	-0.18008	-0.10735	-0.12865	0.02394	-0.025984
GP	-0.06366	0.113268	0.337832	-0.00103	0.12934	-0.10749	-0.09799	0.388692	-0.140264
L	0.066455	-0.18077	0.261757	0.155763	0.212559	0.033621	0.028206	0.119589	-0.050465
mean	1	1	1	1	1	1	1	1	1
Net PK%	-0.09253	-0.01487	-0.17262	-0.16629	-0.1689	-0.05063	0.120866	-0.08356	-0.229108
Net PP%	0.132926	0.087208	0.222997	0.148146	0.315664	-0.14568	-0.05489	-0.10975	-0.10601
OT	-0.16304	0.010458	-0.13503	-0.17971	-0.09047	-0.11876	-0.15402	0.228121	-0.002751
P	-0.09627	0.292429	0.03579	-0.17205	-0.28997	-0.08863	-0.03334	0.026739	-0.028367
P%	-0.08412	0.277413	-0.00614	-0.20527	-0.21145	-0.05031	-0.03014	-0.10808	-0.007759
PK%	-0.09692	-0.10205	-0.29377	-0.15073	-0.10923	-0.03292	0.134049	-0.11504	-0.329391
PP%	0.064714	0.014528	0.219435	-0.18034	-0.12937	-0.13994	0.043255	-0.06378	-0.092687
ROW	-0.0545	0.294647	0.011252	-0.1448	-0.2136	-0.17359	0.02742	-0.04694	-0.077917
RW	-0.0148	0.153991	0.188377	-0.06778	0.011829	-0.14373	0.066911	0.089582	-0.066162
S/O Win	0.082948	0.019838	0.138771	0.305472	0.099151	0.340931	0.00317	-0.02166	0.169425

SA/GP	-0.28264	-0.27923	-0.08498	-0.18963	-0.16676	0.079391	-0.13074	-0.10391	0.03719
Shots/GP	0.126848	0.036513	0.305472	0.073693	0.073693	-0.04216	0.064868	-0.0906	-0.115672
W	-0.02723	0.293238	0.073693	-0.20829	-0.20829	-0.05376	0.028254	-0.05757	-0.025273

Appendix XXXV: Correlation Between Performance and Sentiment for Hashtags



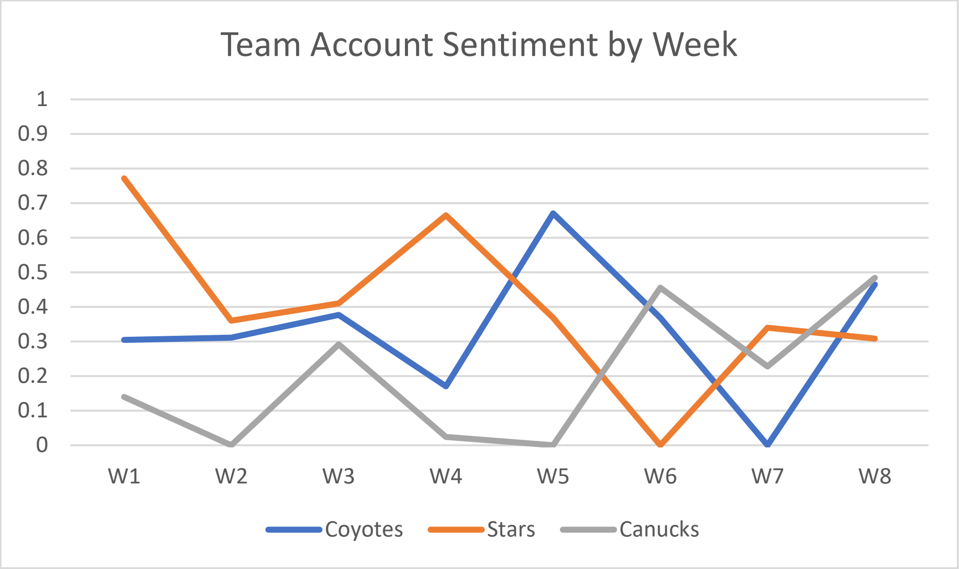
	W1	**W2**	**W3**	**W4**	**W5**	**W6**	**W7**	**W8**	**All**
FOW%	0.479106	0.237032	0.147832	-0.14932	-0.36074	0.31353	0.021986	0.01603	-0.060428
GA	0.02648	0.363875	0.009213	0.001062	-0.15683	0.298615	0.162868	-0.05448	0.037583
GA/GP	0.20921	0.157587	-0.24295	-0.26436	-0.25952	-0.08876	-0.02753	-0.2077	-0.094081
GF	0.315495	0.272201	0.206575	0.022458	-0.20863	0.506601	0.178356	0.194284	0.16601
GF/GP	0.533379	0.094222	0.15391	-0.18147	-0.34253	0.380021	-0.01356	-0.0029	0.088393
GP	0.140302	0.512199	0.301339	0.061408	-0.03552	0.514547	0.207542	0.37747	0.204988
L	-0.14703	0.26592	-0.03867	0.125549	0.041932	0.154492	0.089564	0.227595	0.061307
mean	1	1	1	1	1	1	1	1	1
Net PK%	0.144509	0.201123	0.306515	0.199081	0	-0.13046	0.04366	-0.00498	0.374634

Net PP%	0.12693	-0.11213	0.022125	0	-0.09733	0.260089	0.018339	-0.26138	-0.184086
OT	0.095965	0.173141	-0.04835	-0.22993	-0.03493	0.011071	-0.13848	-0.38046	-0.119808
P	0.328861	0.173763	0.286983	0.014451	-0.05578	0.274549	0.123097	0.162026	0.09013
P%	0.47266	0.047474	0.243088	-0.00889	-0.19157	0.156401	0.06278	0.035329	0.064933
PK%	0.193491	0.189148	0.227664	0.204471	0	-0.13248	-0.0807	-0.02382	0.39675
PP%	0.102171	-0.18079	-0.00402	0	-0.08817	0.242154	-0.10443	-0.24648	-0.19653
ROW	0.325768	0.149512	0.359628	0.104362	0.101692	0.250304	0.131327	0.245653	0.143069
RW	0.453817	0.333889	0.324883	0.109895	0.081426	0.215973	0.077256	0.355494	0.098897
S/O Win	-0.2486	-0.03354	-0.13635	-0.08815	-0.44924	0.03455	0.11161	0.060408	-0.109166
SA/GP	0.519423	0.36506	0.086602	-0.14888	-0.29349	-0.34379	0.06859	-0.15933	0.08039
Shots/GP	0.365182	0.250883	0.125902	-0.20244	-0.45554	0.309733	-0.09801	-0.08262	0.080959
W	0.260611	0.12634	0.291112	0.068671	-0.04643	0.252894	0.167143	0.28422	0.113454

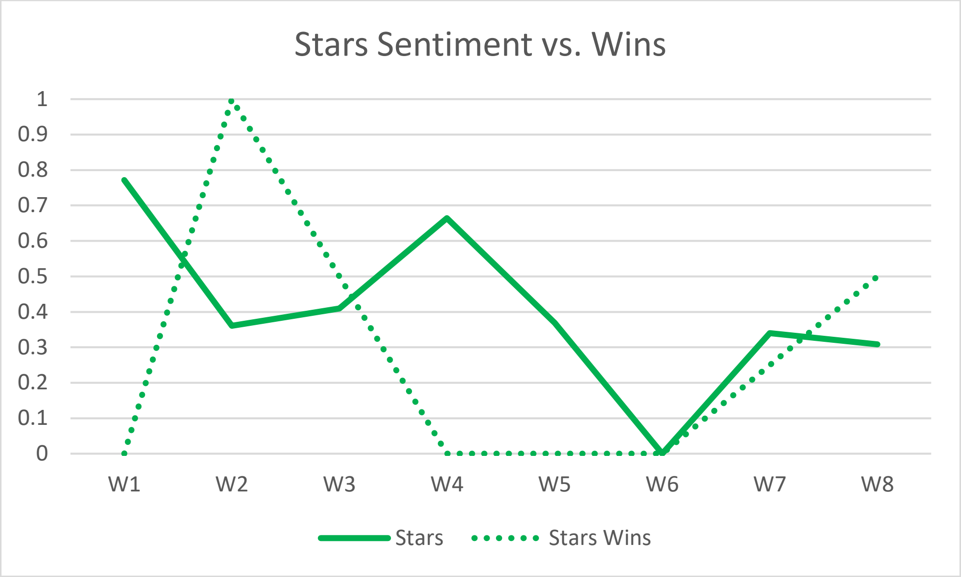
Appendix XXXVI: Most Fluctuating Teams

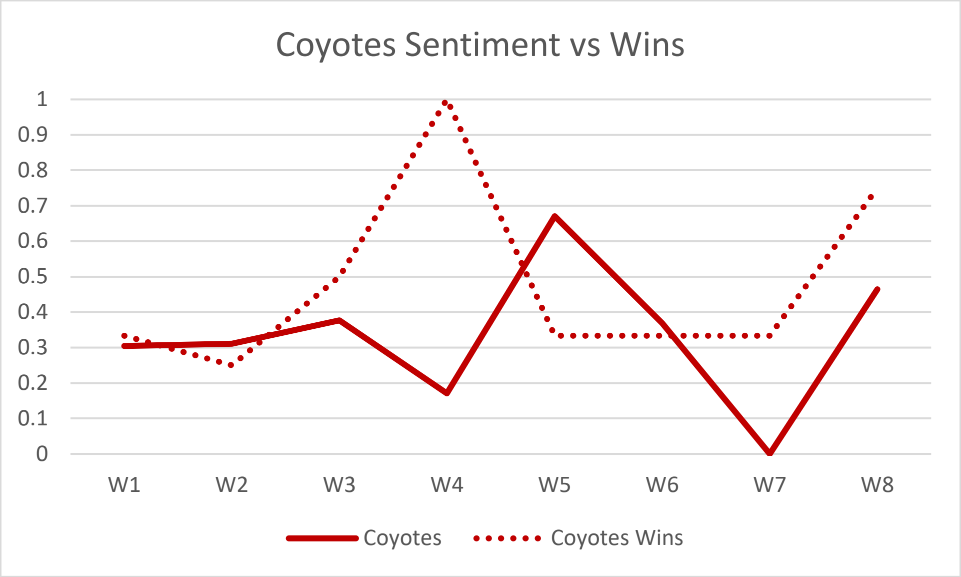
Team	**STD DEV**
Coyotes	0.315374
Stars	0.426551
Canucks	0.276656

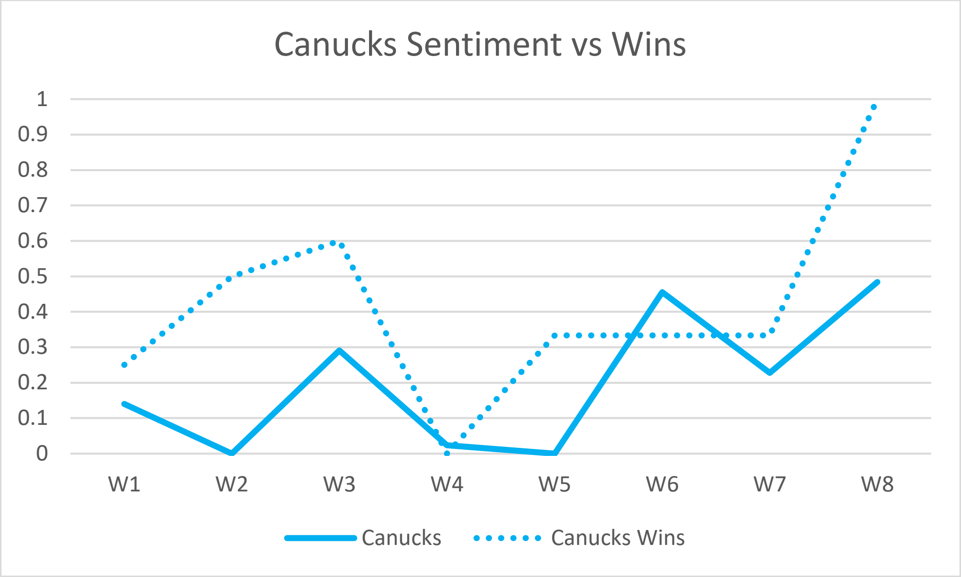
Appendix XXXVII: Most Fluctuating Team Account Sentiment by Week



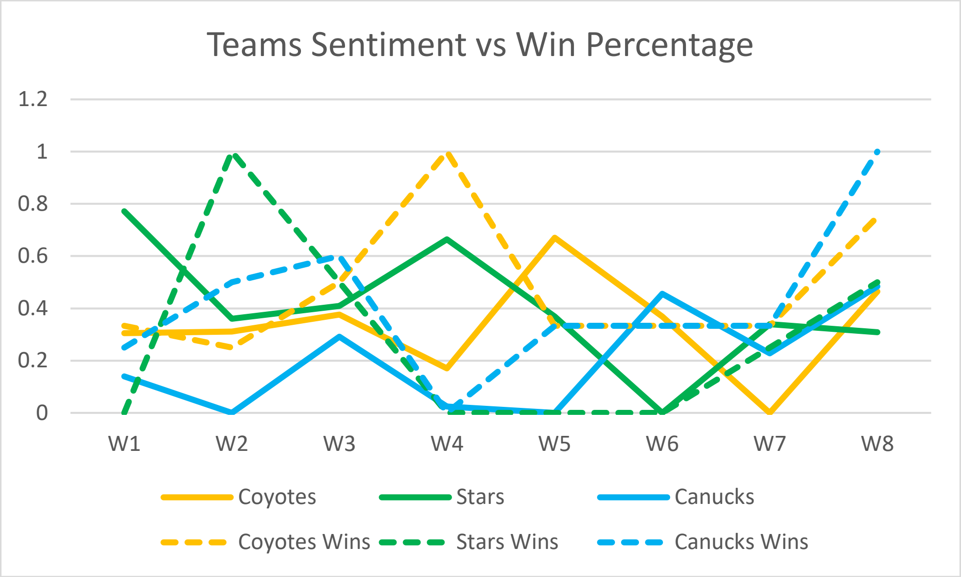
Appendix XXXVIII: Team Sentiment Against Win Percentage Breakdown





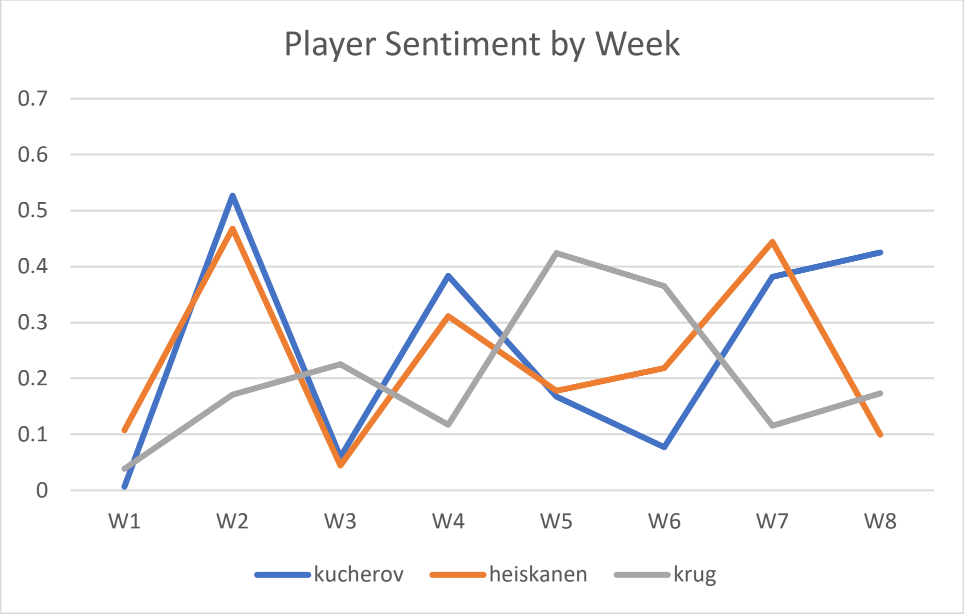


Appendix XXXIX: Team Sentiment Against Win Percentage

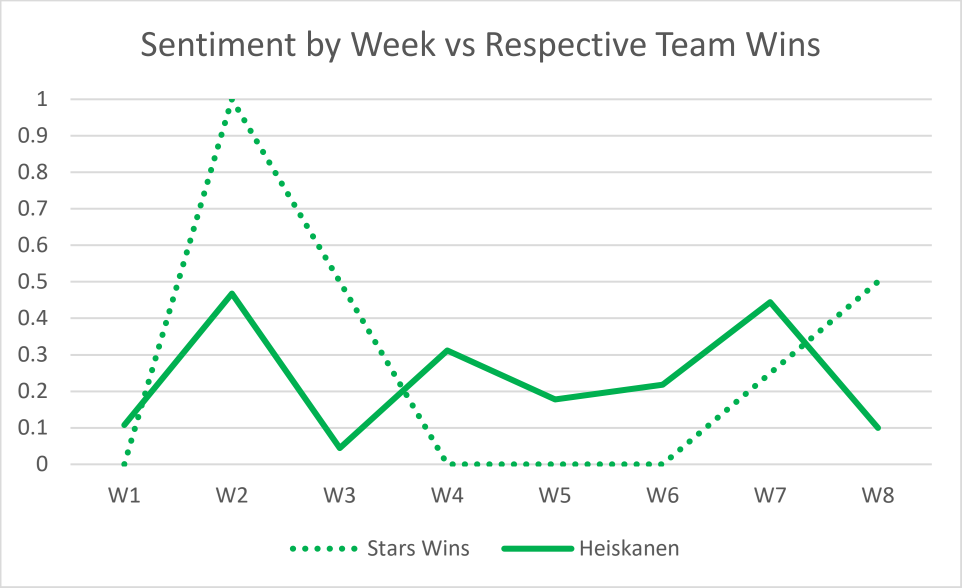


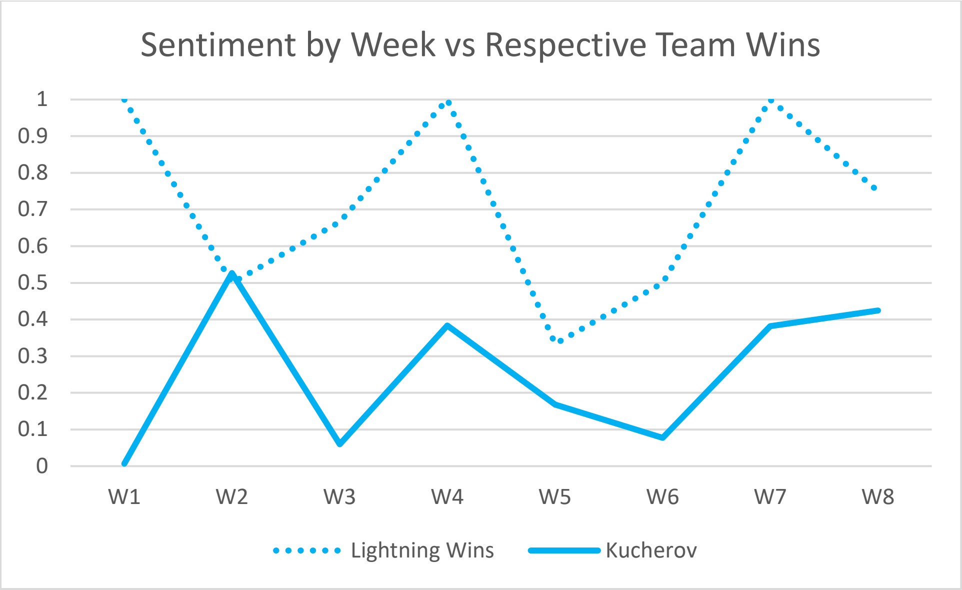
Appendix XL: Highest Fluctuating Players

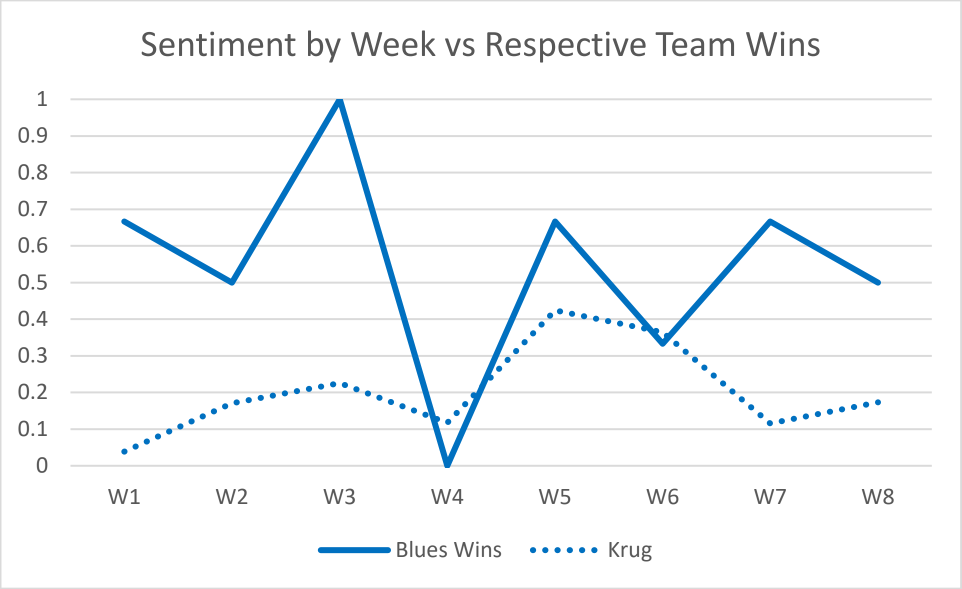
Player	W1	W2	W3	W4	W5	W6	W7	W8	STD DEV
kucherov	0.0067	0.5267	0.059371	0.383175	0.167867	0.077386	0.3818	0.425	0.197944
heiskanen	0.107775	0.467764	0.044467	0.311525	0.1779	0.218333	0.444267	0.099314	0.159578
krug	0.038626	0.170835	0.225486	0.117067	0.424075	0.364837	0.1154	0.173533	0.130683



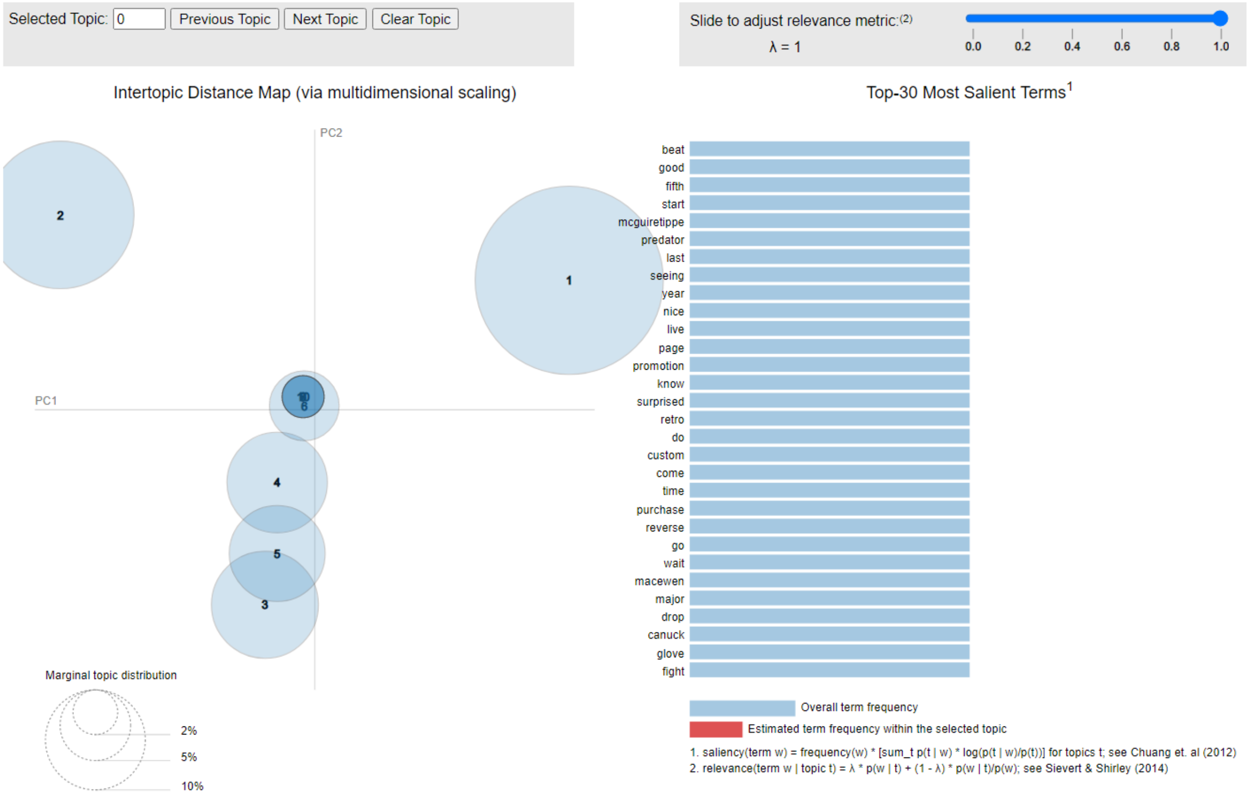
Appendix XLI: Sentiment by Week Against Respective Team Wins


Sentiment by Week vs Respective Team Wins
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
W1
W2
W3
W4
W5
W6
W7
W8
Stars Wins
Heiskanen


Sentiment by Week vs Respective Team Wins
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
W1
W2
W3
W4
W5
W6
W7
W8
Lightning Wins
Kucherov

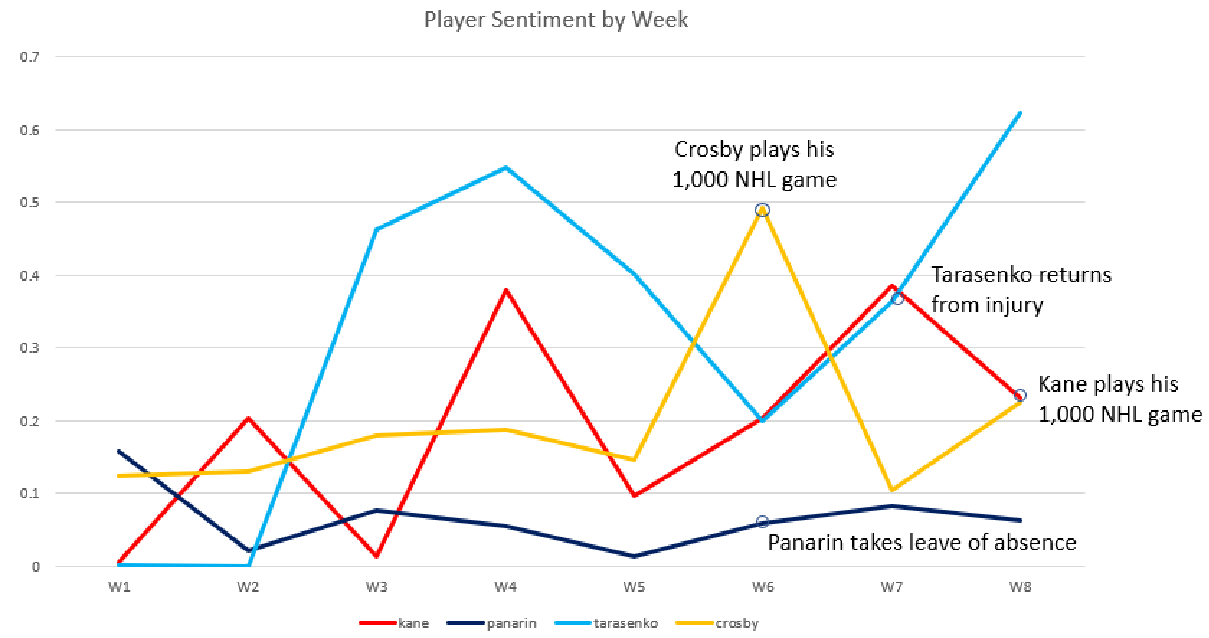


Appendix XLII: Latent Dirichlet Allocation (LDA) Results

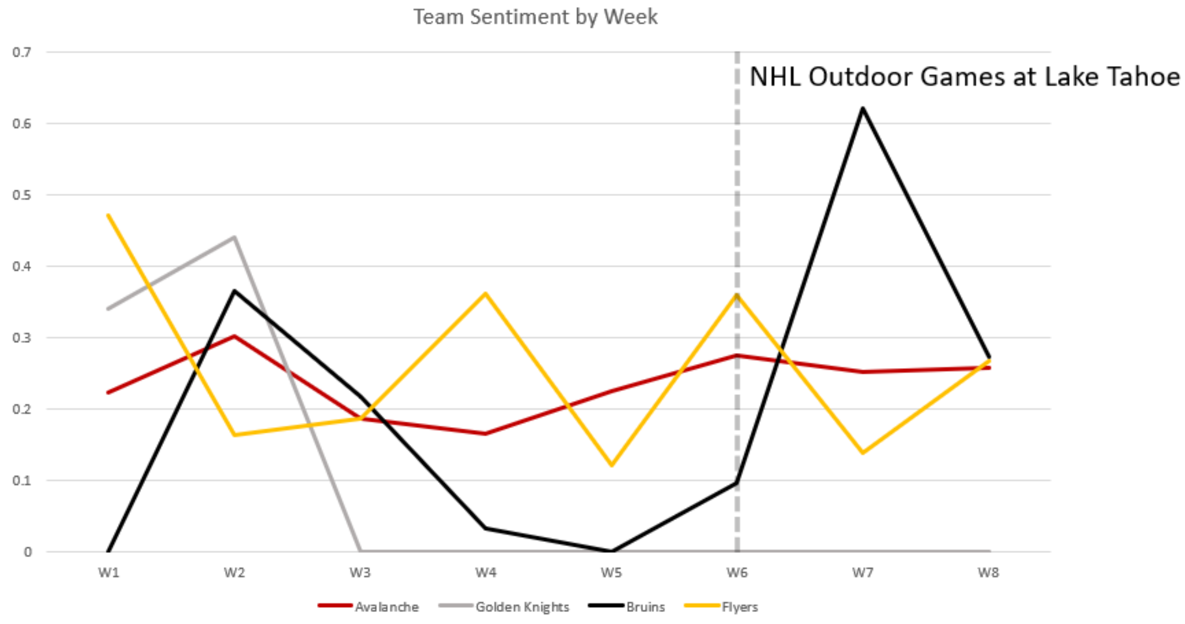


1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

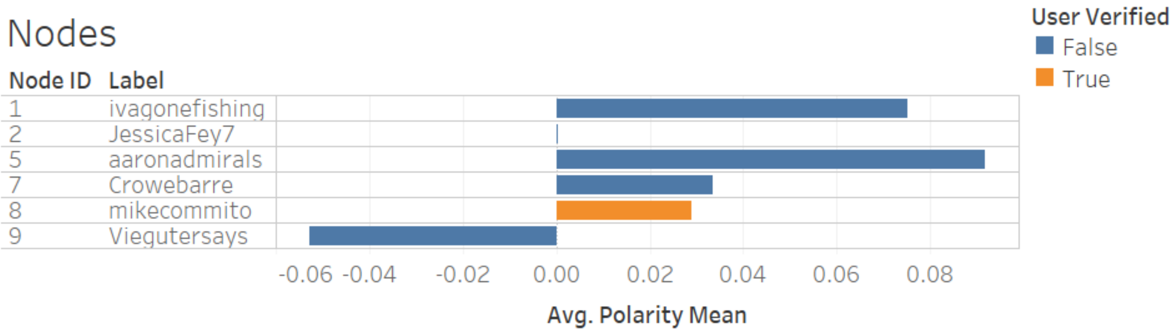
Appendix XLIII: Player Sentiment by Week



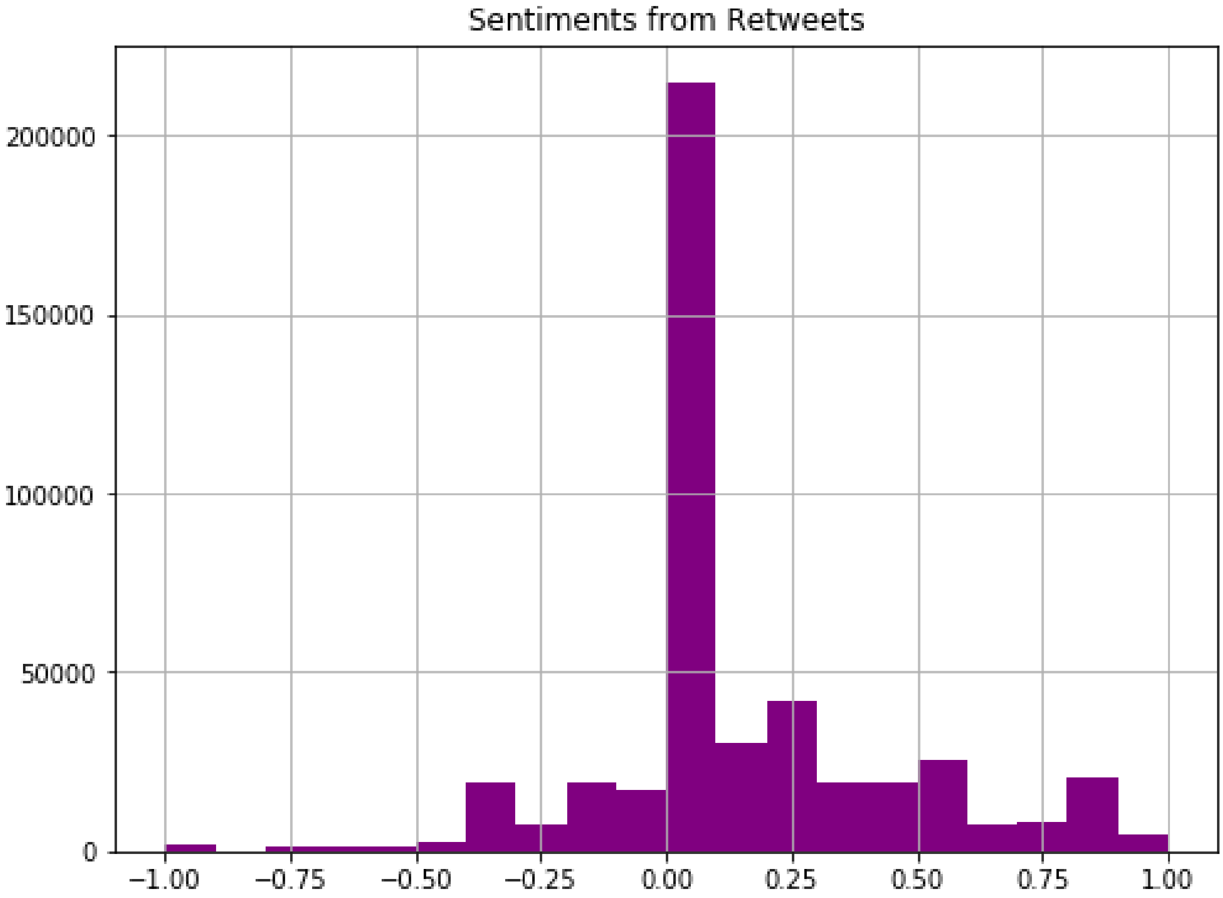
Appendix XLIV: Team Sentiment by Week



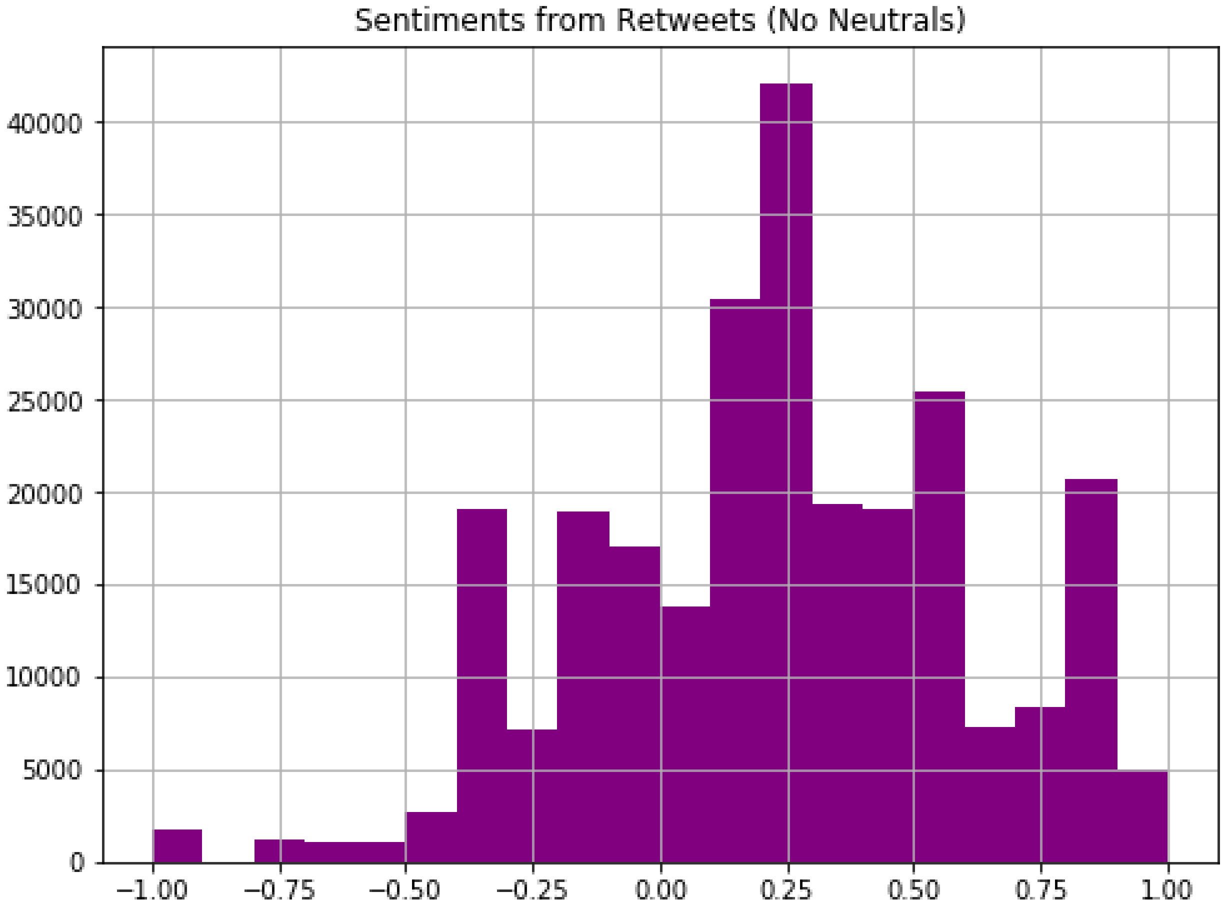
Appendix XLV: Leading Accounts in Each Node



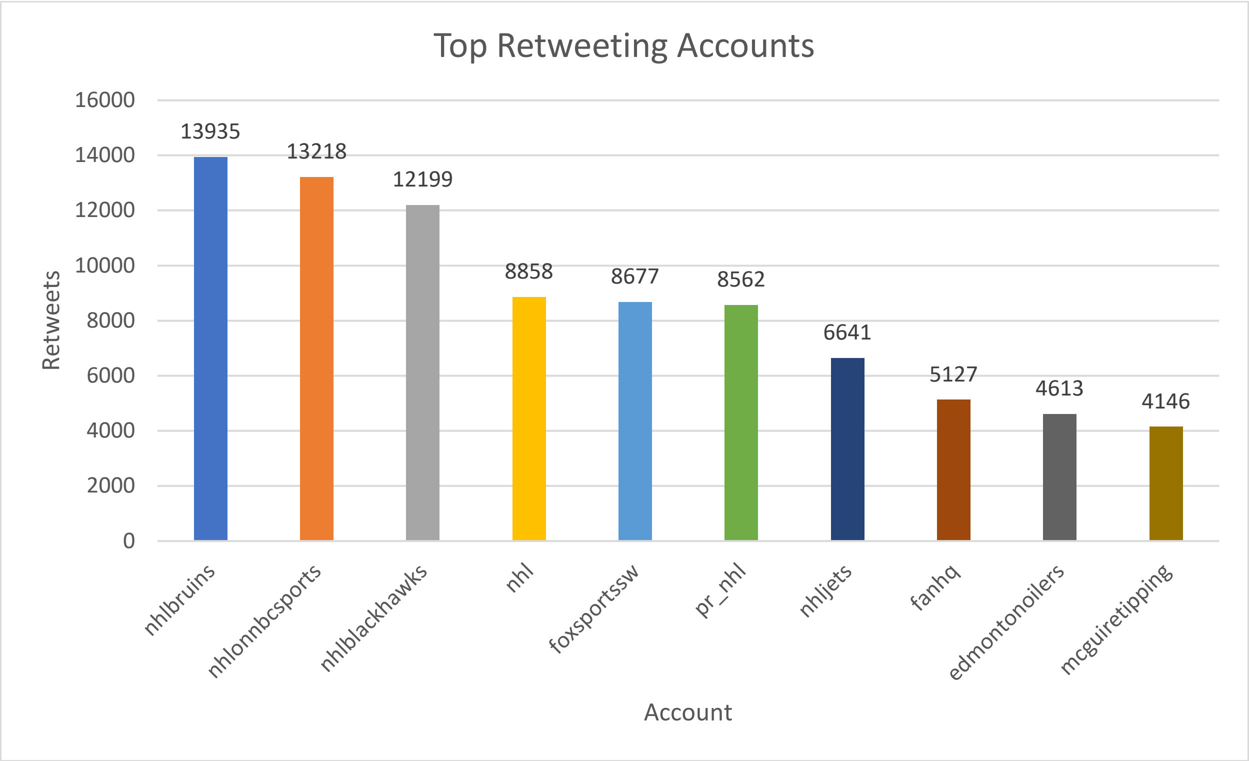
Appendix XLVI: Sentiments from Retweets with Neutrals



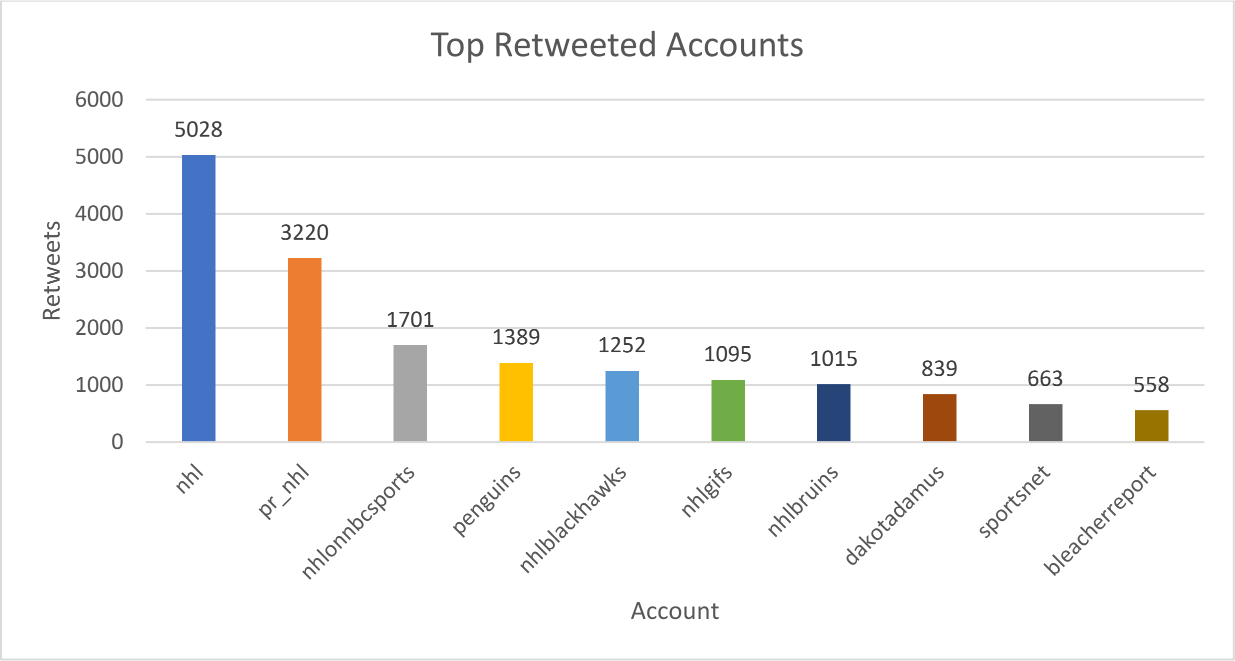
Appendix XLVII: Sentiments from Retweets without Neutrals



Appendix XLVIII: Top Retweeting Accounts



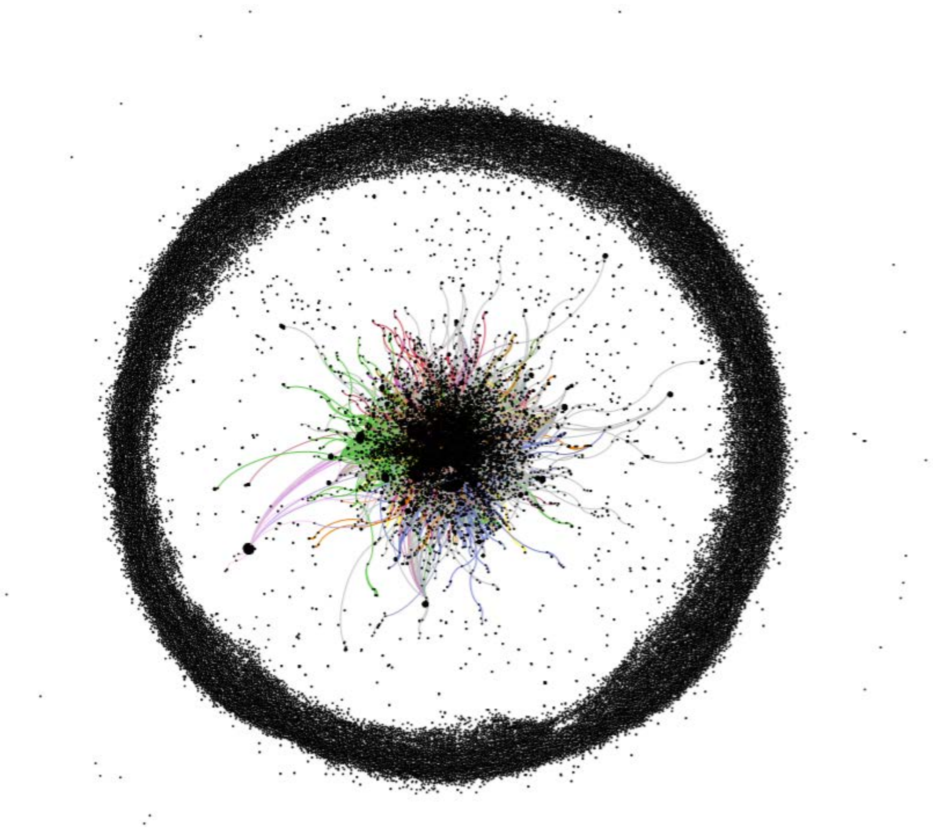
Appendix XLIX: Top Retweeted Accounts



Appendix L: Retweeted Accounts

Account	# Times this account retweeted someone else	Who Retweeted?	# Times Retweeted
nhl	0	5028	8858
pr_nhl	0	3220	8562
nhlonnbcsports	0	1701	13218
penguins	77	1389	1521
nhlblackhawks	0	1252	12199
nhlgifs	0	1095	1183
nhlbruins	0	1015	13935
dakotadamus	0	839	1024
sportsnet	0	663	1195
bleacherreport	0	558	752

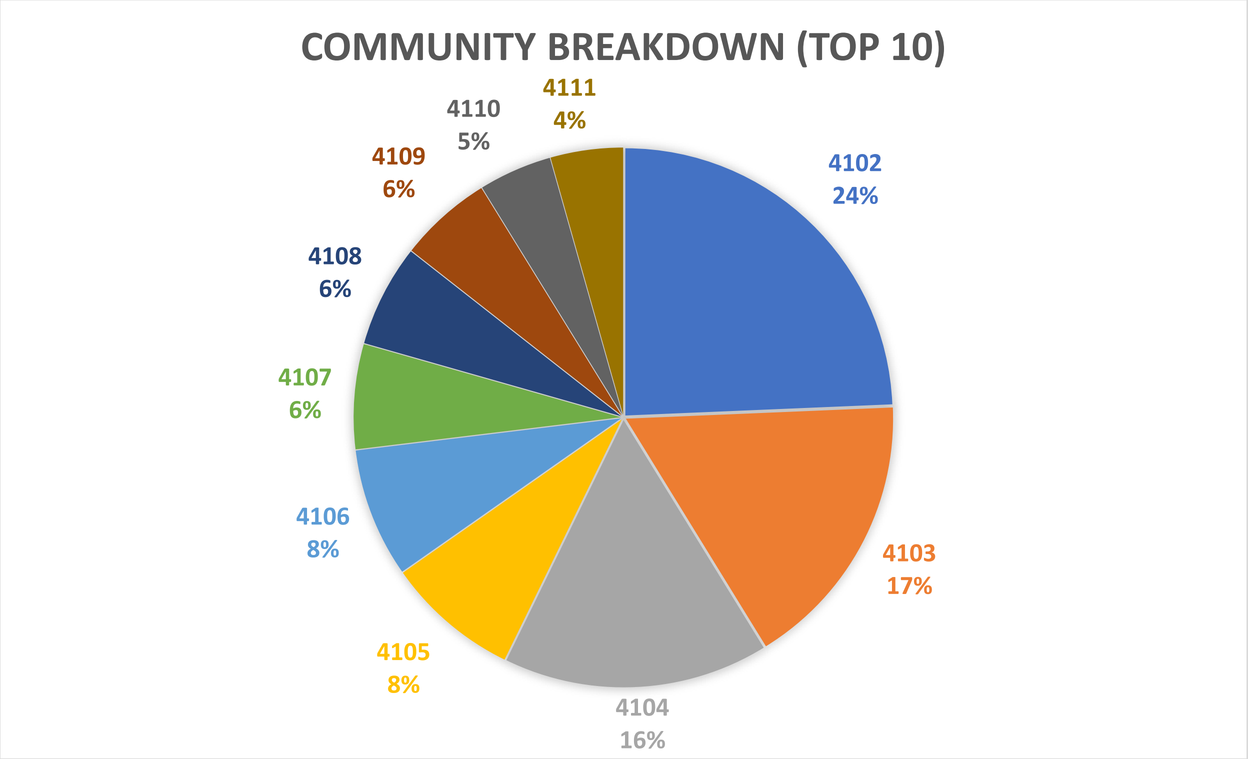
Appendix LI: Social Network Results



Appendix LII: Node Breakdown

Community Number	% Nodes
4102	6.11%
4103	4.24%
4104	4.02%
4105	2.02%
4106	1.96%
4107	1.59%
4108	1.55%
4109	1.41%
4110	1.11%
4111	1.10%

Appendix LIII: Top Ten Node Breakdown



Appendix LIV: Top Prestige Scores

Account	Community	Prestige Score
joelthesakic	4102	1
NHLSabresNews	4102	0.97429753
InglouriousBee	6717	0.958539798
TBL_Hockey	5599	0.95488189
flyer4life	38411	0.938936385
Legend_of_Lando	4102	0.844053657
apdodds_	4102	0.822048468
DBrysh	4102	0.822048468
dbanns	4102	0.822048468
KabilanLingam	38411	0.821203183

Appendix LV: Team Fanbases

Rank	NHL Team	Fans	Stanley Cups
1	Chicago Blackhawks	2,867,678	6
2	Boston Bruins	2,211,511	6
3	Pittsburgh Penguins	2,062,579	5
4	Detroit Red Wings	2,029,284	11
5	Montreal Canadiens	1,640,600	24
6	New York Rangers	1,537,175	4
7	Toronto Maple Leafs	1,380,450	13
8	Philadelphia Flyers	1,179,137	2
9	Vancouver Canucks	1,033,870	0
10	L.A Kings	971,448	2
11	San Jose Sharks	956,423	0
12	Washington Capitals	804,263	1
13	Colorado Avalanche	762,986	2
14	St. Louis Blues	745,089	1
15	Minnesota Wild	651,986	0
16	Edmonton Oilers	580,124	5
17	Tampa Bay Lightning	564,299	1
18	Buffalo Sabres	503,525	0
19	Dallas Stars	483,601	1
20	New Jersey Devils	477,853	3
21	Anaheim Ducks	424,352	1
22	Nashville Predators	409,899	0
23	Winnipeg Jets	394,482	0
24	Calgary Flames	384,363	1
25	Vegas Golden Knights	349,888	0
26	Ottawa Senators	335,522	0
27	Columbus Blue Jackets	317,671	0
28	New York Islanders	304,074	4

29	Arizona Coyotes	300,605	0
30	Carolina Hurricanes	285,188	1
31	Florida Panthers	197,721	0

REFERENCES

Academy, U.S. Sports. “The Impact of Social Media in Sports.” The Sport Digest, 27 Nov. 2018, thesportdigest.com/2018/11/the-impact-of-social-media-in-sports/.

“Athletes and Twitter: Using Sentiment Analysis to Improve Athletic Performance.” UC Institute for Prediction Technology, predictiontechnology.ucla.edu/athletes-and-twitter-using-sentiment-analysis-to-improve-athletic-performance/.

Bruns, Axel, Weller, Katrin, & Harrington, Stephen (2014) Twitter and sports: Football fandom in emerging and established markets. In Bruns, A, Mahrt, M, Weller, K, Burgess, J, & Puschmann, C (Eds.) Twitter and society [Digital Formations, Volume 89]: Peter Lang Publishing, United States of America.

Daren, Sarah. “The Pros and Cons of Athletes Using Social Media, Coach's Clipboard Basketball Coaching.” Coach's Clipboard Basketball Coaching, Coach's Clipboard, 23 Mar. 2018, www.coachesclipboard.net/athletes-and-social-media.html.

“Does the Media Impact Athletic Performance?” The Sport Journal, 18 Apr. 2017, thesportjournal.org/article/does-the-media-impact-athletic-performance/.

Exsci. “Social Media Negatively Impacts Sports Performance.” Online Exercise Science Degree, 15 Mar. 2019, exsci.cuchicago.edu/study-shows-late-night-social-media-use-may-decrease-athletic-performance/.

Forrest, D., & Simmons, R. (2002). Outcome uncertainty and attendance demand in sport: The case of English soccer. The Statistician.

Forrester, Nicole W. “The Selfie Olympics: What's the Impact of Social Media on Performance?” The Conversation, 9 Feb. 2020, theconversation.com/the-selfie-olympics-whats-the-impact-of-social-media-on-performance-92273.

"How Media Use Hurts Athletes." Psychology Today, Sussex Publishers, www.psychologytoday.com/us/blog/the-power-prime/201701/how-media-use-hurts-athletes.

John Price, Neil Farrington & Lee Hall (2013) Changing the game? The impact of Twitter on relationships between football clubs, supporters and the sports media, Soccer & Society, DOI: 10.1080/14660970.2013.810431.

Kapanipathi P., Jain P., Venkataramani C., Sheth A. (2014) User Interests Identification on Twitter Using a Hierarchical Knowledge Base. In: Presutti V., d'Amato C., Gandon F., d'Aquin M., Staab S., Tordai A. (eds) The Semantic Web: Trends and Challenges. ESWC 2014. Lecture Notes in Computer Science, vol 8465. Springer, Cham.

Koch, Connor. "Analyzing Yankees and Red Sox Sentiment Over the Course of a Season." *Bryant University*, 2020.

Melendez, Steven. "The Impact of Social Media on Athletes." PantherNOW, 22 Feb. 2018, panthernow.com/2017/10/17/the-impact-of-social-media-on-athletes/.

NHL, www.nhl.com/stats/

Paul, R., & Weinbach, P. (2007). The uncertainty of outcome and scoring effects on Nielsen ratings for Monday Night Football. Journal of Economics and Business.

Pegoraro, Ann. "Look Who's Talking—Athletes on Twitter: A Case Study". International Journal of Sport Communication. <https://doi.org/10.1123/ijsc.3.4.501>. Web. 17 Apr. 2020.

Severini, Thomas A. *Analytic Methods in Sports*. CRC Press, 2020.

"The Impact of Social Media on the Sports Industry." PRLab StudentStaffed Public Relations Agency the Impact of Social Media on the Sports Industry Comments, 1 Jan. 1969, www.bu.edu/prlab/2018/10/29/the-impact-of-social-media-on-the-sports-industry/.

Thomas A. Severini, 2015, taseverini.com/index.html.

Tweepy, www.tweepy.org.

Watanabe, Nicholas, et al. “Major League Baseball and Twitter Usage: The Economics of Social Media Use.” Journal of Sport Management, vol. 29, no. 6, Nov. 2015, pp. 619–632. EBSCOhost, doi:10.1123/JSM.2014-0229.

Witkemper, Chad & Lim, C.H. & Waldburger, A. (2012). Social media and sports marketing: Examining the motivations and constraints of Twitter users. Sport Marketing Quarterly.

Yang Yu, Xiao Wang, World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans’ tweets, Computers in Human Behavior, Volume 48, 2015, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2015.01.075.