NEW YORK ISLANDERS - USING ANALYTICS TO GUIDE SOCIAL MEDIA STRATEGY

NEW YORK ISLANDERS – SPRING 2020

Jack Behrens
Sloan School of Management
jackberh@mit.edu

Kevin Ma
Sloan School of Management
kevinsma@mit.edu

Henry Martin
Undergraduate MIT Student
hemartin@mit.edu

Adedayo Aderibole
Department of Electrical Engineering and Computer Science
adedayo@mit.edu

Ferran Vidal-Codina
MIT Sports Lab
fvidal@mit.edu

Christina Chase *
MIT Sports Lab
cchase@mit.edu

Anette E. Hosoi †
MIT Sports Lab
peko@mit.edu

May 1, 2020

ABSTRACT

The abstract should summarize the purpose and outcomes of the report in 150 words or less. This typically involves declaring the main take away from the report,[3] outlining the problem, describing what methodology was implemented and why, and giving a short summary of the results. This section is typically written last after everything else is in place. Try to write as clear as possible.

1 Introduction

Our problem is an open-ended one: how can we help the New York Islanders understand their Twitter followers, and provide insights to help the organization better engage fans on this platform? Currently most professional sports organizations, including the Islanders, view their social media as a black box and are not maximizing the potential of their twitter presence. This is an extremely relevant problem for the teams, specifically the Islanders in this project, as they could better use social media to help run a successful organization financially and draw in more fans.

The main technical challenge is collecting and analyzing data to gain meaningful insights that the Islanders organization can use to make business decisions. We accomplished this by splitting the project into two separate prongs, each with their own separate goal. One goal was to cluster the followers on the Islanders official account to better understand who the fans are. The other is look at all activity from the Islanders accounts and activity with hashtags and keywords relating to the Islanders to understand how the fans interact with one another and the official team account.

The main data we have used is Twitter data, which we will discuss more in the methodology section below, along with how we approached the problem and worked with the data. In the results section we will discuss our clusters and the insights from the other prong of the project, along with what they mean for the islanders organization. In the conclusion and future work section we will discuss our takeaways from the project and potential future projects for the Islanders Organization or future 2.980 students.

*Corresponding author
†Corresponding author

[3] "The main take away from the report" is usually the topic sentence of the abstract, e.g. "By analyzing/building 'name' we were able to (better) achieve/model the 'desirable outcome'."

2 Methodology

Our major resource is a trove of data obtained through Twitter's API using a Twint github package. We are collecting information on tweets, interactions, and profiles of accounts that tweet certain keywords or follow the official @NYIslanders account. We are also partnering with Sam's team to get their insights into things like: comparable Islanders-focused Twitter accounts, reasonable ways to categorize tweet content, and expanded list of keywords and hashtags. As far as tools go, we are storing our data on BigQuery, and using Google cloud tools such as CoLab, Sheets, and Docs, as well as Slack to collaborate on our various workflows. Here are the three different datasets we were mainly focusing on in our project.

- All Islanders-related tweets from the 2018-2019 and 2019-2020 seasons, targeting specific hashtags and keywords.
- All official @NYIslanders tweets from the 2018-2019 and 2019-2020 seasons
- All tweets from September 2019 - April 2020 for users who used 'isles' more than once during that time

Here we have several subsections, each is talking about how we approached the different prongs of our approach in terms of methodology, the data breakdown above will apply to both prongs.

2.1 Clustering Analysis

The process of separating the New York Islanders twitter fans involved two major steps, Feature Engineering and Clustering. These steps ensured the segmentation of fans into clear and interpretable clusters.

2.1.1 Feature Engineering

This is a process of selecting and converting raw data into a form that conveys fundamental information about the problem that is being solved. Feature engineering was applied to transform raw data of twitter users that interacted with New York Islanders related content to features/attributes that was fed into a set of machine learning algorithms.

Raw features employed in our algorithms can be classified as either numerical or binary. Numerical features include number of tweets mentioning the islanders, number of followers, total historical tweets, etc. While logical variables such as the presence of user background image or islanders related content in user bio were classified as binary features. The full feature list directly obtained from the database can be found in the Appendix. Afterwards, the numerical features with large statistical ranges were converted to a logarithmic scale and then standardized to have a mean of 0 and standard deviation 1 to improve the performance of the machine learning algorithms. In addition, the binary features were encoded with the one-hot scheme.

Further feature selection and reduction algorithms such as random forest and principal component analysis (PCA) were performed on the standardized feature space to select the most important features and reduce the dimensionality of the feature space. The resulting data was then fed input to various clustering algorithms.

2.1.2 Clustering

In this subsection, the procedure employed to group twitter users that interact with the New York Islanders content will be explained. The results of the clustering analysis will enable properly ranking the studied twitter users on the New York Islanders fan ladder. In this work, the Kmeans and Hierarchical techniques were employed to for this purpose. The clustering approach involved inputting the feature data into the algorithms and selecting the optimal number of clusters. The optimal number of clusters was chosen based on the silhouette score. This score provides information on the proximity of observations within a single cluster and observations in different clusters.

The Kmeans algorithm works by initially placing K centroids randomly within the dataset. Each observation (twitter user) is placed in a cluster to minimize its distance to its cluster cemtroid. Afterwards, a new centroid of each K cluster is moved to become the mean of all observations in that cluster. This process is repeated until convergence is reached. Hierarchical clustering can be classified as either Agglomerative or Divisive clustering. Agglomerative (Bottom-to-top approach) clustering begins by treating each twitter user as a separate cluster, then identifies and merges the two most similar clusters. This iterative process continues until all the clusters have merged together. Divisive (top-to-bottom) clustering starts with every observation in a single cluster and continuously breaks the cluster down until all observations are in separate clusters. The aim of hierarchical clustering is to build a hierarchy of clusters. The two most important factors that impact the performance of hierarchical clustering are the linkage criteria and measure of distance. In this work, we applied Kmeans, Agglomerative and Divisive clustering to segment twitter users into different New York Islanders fan category. The results of the clustering algorithm will be presented in the Results and Discussion section.

2.2 Data Analysis

Using the data from above, we went through a data cleaning process in Colab notebooks, creating new Csv files to do analysis with. We created two main avenues for analysis, looking at the tweets sent from the Islanders account and then the tweets from different fans. For both sides of the interaction we wanted to answer the same question, how are fans interacting with the Islanders on twitter, either through tweeting keywords and hashtags, or through the Islanders official account.

For both of these we segmented the tweets themselves, breaking down as much detail as possible and then adding in descriptions. For example, we broke down day of the week, whether it was a gameday or not, what the result of the game was and several other factors. This gives us a chance to break out fan activity into different buckets and see how the activity changes as a whole.

For the Islanders tweets from the official account, we focused on breaking down the type of tweet as well, whether it had media, hashtags, mentions, or a link to the Islanders website. From this breakdown we focused on how these different factors effect overall performance of the tweet with new metrics we created.

Following industry insight, we wanted to create metrics to help compare the different actions a user can take on a tweet, a RT, favorite or reply. We created three metrics, raw score, reach score and depth score. The formulas for those three are below.

$$\begin{aligned} Raw\ Score &= RTs + Replies + Favorites \\ Reach\ Score &= (1.0 * RTs) + (0.25 * Replies) + (0.1 * Favorites) \\ Depth\ Score &= (0.25 * RTs) + (1.0 * Replies) + (0.1 * Favorites) \end{aligned} \tag{1}$$

We used these metrics instead of just RT's or replies to get a better sense of the impact and interaction with the Islanders account.

3 Results and Discussion

For the results and discussion section we will once again break it down to our two subsections of the project and discuss separately before giving conclusions and takeaways more generally.

3.1 Clustering

The results of our clustering can be talked about best with the following 2 figures and then summarized from there.

Figure 1 shows the silhouette scores corresponding iwth the number of clusters for a variety of different clustering algorithms. This has given us a way to choose the best possible clustering algorithm.

Figure 2 is a visualization of what our clusters look like. This tsne plot is a plot with the features reduced to just two dimensions to show the clusters and how they are grouped. As you can see, the outside larger clusters are all pretty close together and well grouped, with some noise in the middle.

In addition to these two plots, the final version will have a longer discussion of our clusters and the takeaways from them, along with more visualizations on the clusters themselves.

3.2 Data Analysis

Here are some of the most interesting and relevant visualizations and discoveries of our data.[4]

For our results in the data analysis section, we will show the most impactful images and then the takeaways and actionable insights from these graphs.

Figures 3 and 4 come from analyzing the tweets from the Islanders Account itself while the data from 5 is the general activity about Islanders data set. Figure 6 looks at both comparing the lockdown period of hockey activity.

The takeaway from figure 3 is that the level of interaction on Islanders tweets are much higher for the postgame period if the team won the game. This means that more people are looking at the account in general and sharing the content, meaning it could be a good time for promotion, ticket sales or other marketing content.

[4]We will add more to this section in the final draft, consider these just a sample

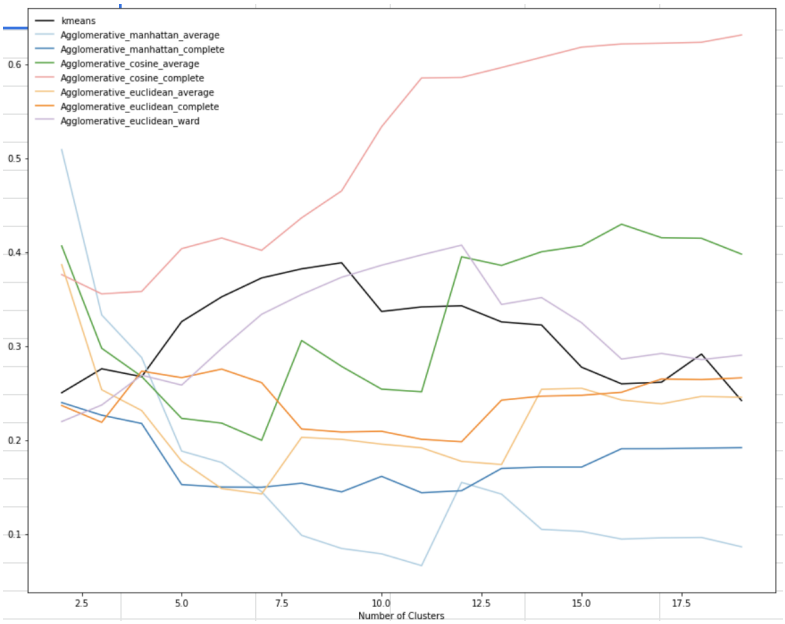

Figure 1: Silhouette Score graph

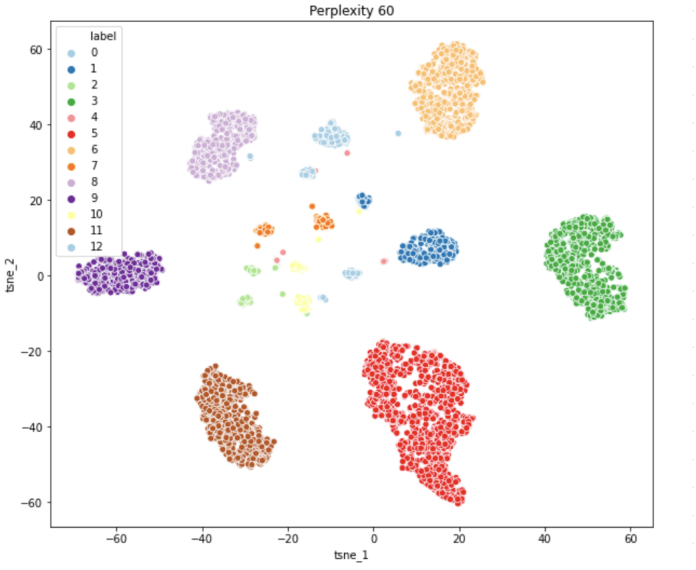

Figure 2: Clustering Visualization

The takeaway from figure 4 is that twitter content from the Islanders is much more widely shared and consumed when it includes media, which is constant whether it is a gameday or not. In this case media is an embedded picture, video, or gif. I would recommend adding in media to any post that the Islanders think is important or they want to spread.

Figure 5 is encapsulating how the fans use twitter depending on the result and location of the game. As we can see, the most interacted with type of game is a home loss, when there is the most activity from a fan using the keywords and hashtags. This is interesting as it not mirrored in the islanders content, suggesting a difference in the way people tweet on their own account versus their interaction with the Islanders account.

The takeaway from figure 6 is that there may be some differences in the Islanders account activity and the way people are tweeting about the Islanders post COVID-19 and there may be some way to fix it.

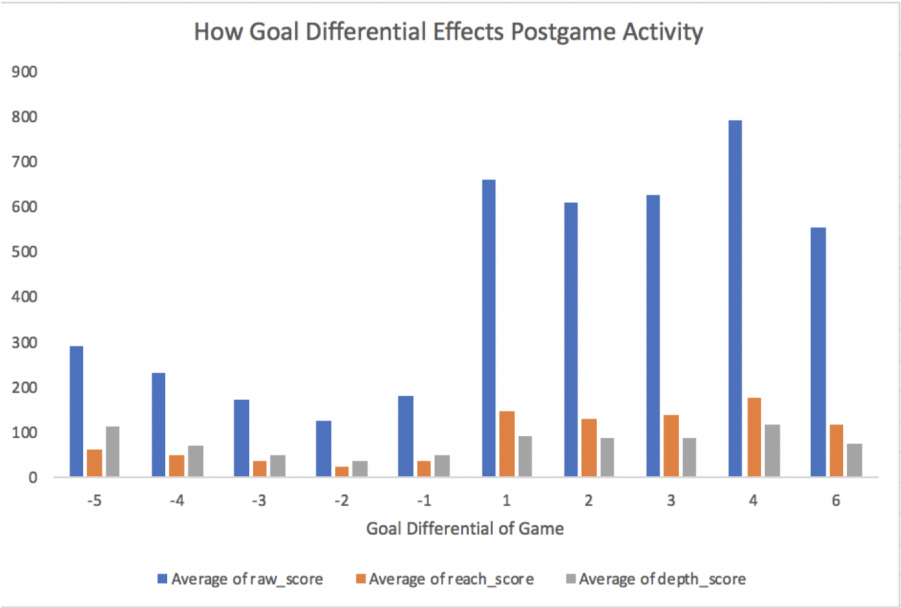

Figure 3: Postgame Activity effect based on game result

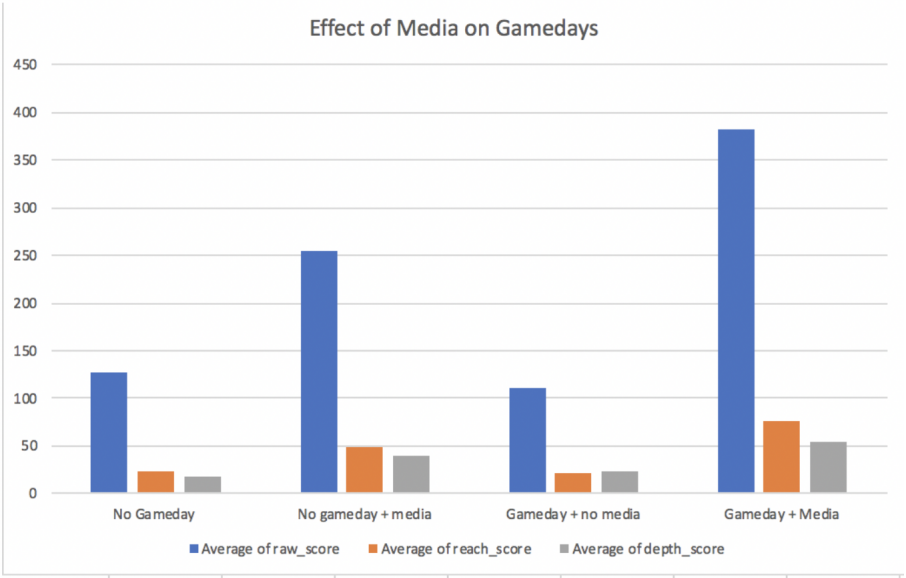

Figure 4: Graphic showing how media effects response to offical account tweets

4 Conclusions and future work

This section will be done in the final proposal, once we have talked through this as a team in the coming days. This will include future projects for the Islanders, our biggest takeaways and some of the areas we thought were most interesting from a technical perspective and from a business side for the organization.

References

A Appendix

Your appendix information goes here

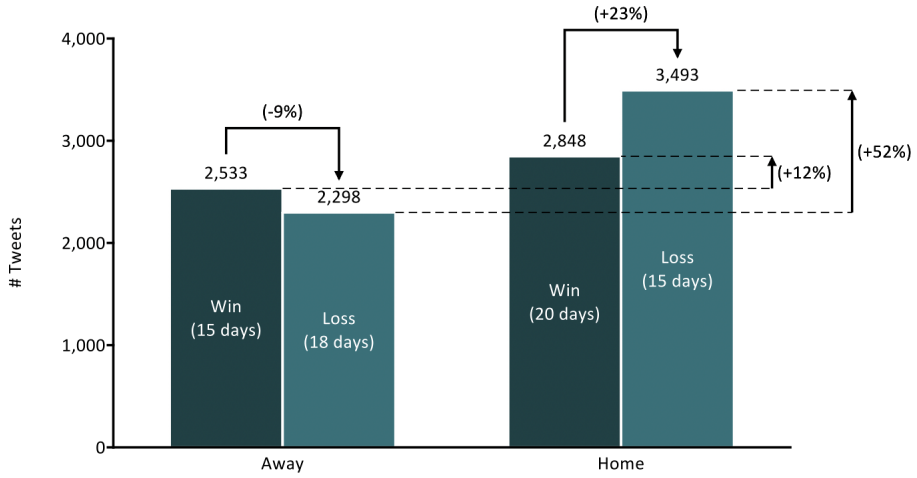

Figure 5: Home away effect on Islanders twitter activity

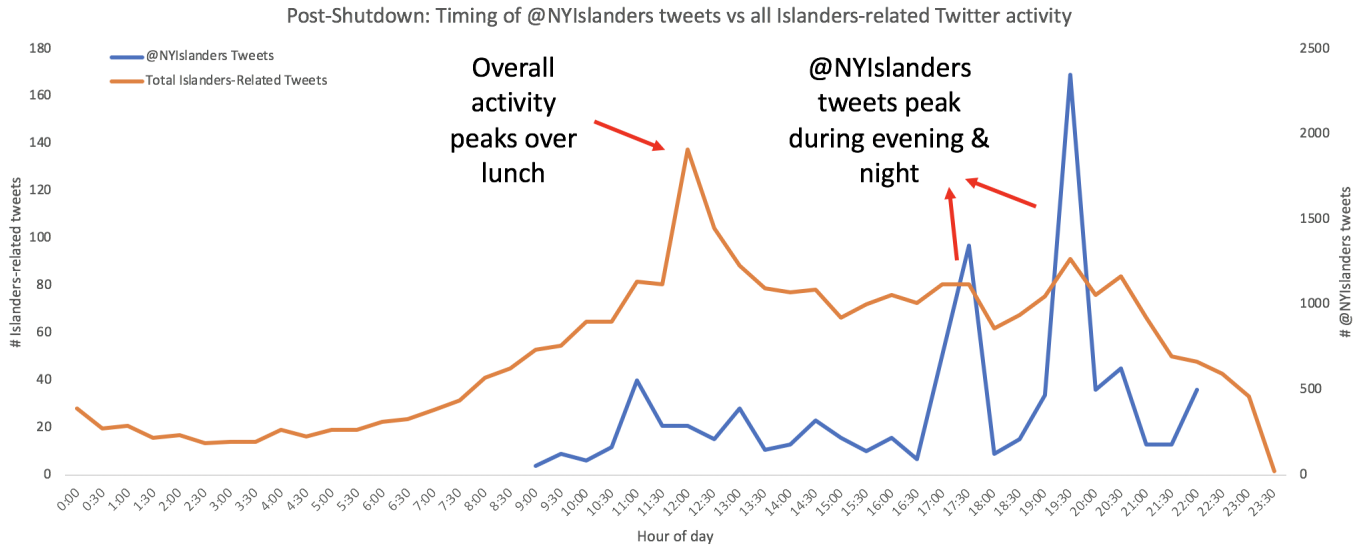

Figure 6: Difference of peaks in activity between Islanders accounts and Islanders related tweets