Springboard DSC--Capstone Project 1
NFL Breaking News Classifier
By Logan Larson
May, 2019

Problem Statement: Since there doesn't exist an application that can deliver comprehensive NFL breaking news updates in real time, the goal of this project is to ultimately develop a system for classifying relevant tweets from a Twitter feed of NFL reporters to push the appropriate information to users within seconds of posting.

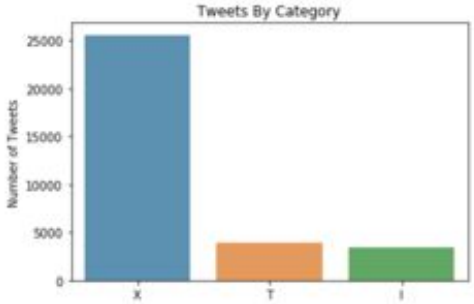
The Client: This is a problem for both my employer and those who participate in fantasy football. For my employer, it's becoming increasingly difficult to compete with Matthew Berry's Fantasy Life app, which currently provides mobile alerts with a slight delay (though it only covers the highest-priority news, not everything that could be relevant to veteran users). For fantasy football participants, having instant knowledge of all relevant transactions could potentially provide a significant advantage over the competition, particularly in terms of being the first to add free agents from the waiver wire. In a broader sense, participants otherwise must go into their mobile apps to refresh news streams to get updates on the non-highest-profile players. If they don't, they likely miss around half of actionable news that could have otherwise improved the position of their team.

The Data: In the NFL world, there's conveniently one incredibly well-networked and accurate reporter, Adam Schefter, who breaks a large percentage of the high-profile breaking news. Convenient as well, Schefter exclusively uses Twitter to break news, so we can collect the data we need by scraping his timeline. Working with the Twitter API to do so, I ultimately collected a sample of over 30,000 of his tweets spanning from 2010 to 2019. Steps I took to clean my data involved deleting irrelevant features, converting dates to datetime objects, and writing functions to clean up URLs and hashtags.

Initial Findings:

An initial goal of my project is to take Schefter's tweets and categorize them into three categories: X, T and I. The 'T' label designates a tweet is transactional or, in other words, any tweet that contains information about any player transactions that impact their contract status or playing eligibility. The 'I' label designates any workload impediments such as injury updates or role changes. The 'X' label is any and everything else.

We care about 'T' tweets because we need to know which players are on which teams -- and within those teams, which of the players are eligible to compete in the first place. Then we get more specific with the 'I' tweets to gain further information on which eligible players are healthy to play and, subsequently, how much playing time they can be expected to receive. The 'X' tweets are then entirely irrelevant for our purposes since they don't offer us any further, actionable information. For the purposes of fantasy football, the 'I' tweets are generally most important to us because the most relevant players don't often change teams, but the injury bug doesn't discriminate.



After getting a sense for which words are most characteristic of each class of tweet, we next looked at the makeup of each class independent of the underlying words. A few observations can be made:

1. 'T' tweets (15.51 words, 97.39 characters) tend to be significantly shorter than any 'I' (18.18, 110.22) and X (18.09, 112.77) with a significance of r = 0.33.
2. All relevant tweets ('T' and 'I') have at least three words but can still take up the entire 140-character max that Twitter allows.
3. The presence of a hashtag is 2.5 times more likely in 'X' tweets than either 'I' or 'T' tweets. Likewise, the presence of an @ sign is 7 times more likely.

Next Steps:

Next I intend to apply different advanced NLP data cleaning techniques such as lemmatization and ultimately apply different techniques to find the optimal classifier for our data.