Charlie Hill
Professor Rivas
Artificial Intelligence
12 December 2016

Final Write-up

Abstract

For many people, flying can be a nightmare. Whether it be a fear of flying, a crying baby or delays, flying can easily be an unenjoyable experience. Whenever you ask someone about their flight most of the time they will find something to complain about. With the internet, everyone is able to vocalize their opinion on their own personal experience. Many people vocalize their opinions through social media. One of the most popular social media sites used today is Twitter. Using the information that comes from how users react to their airline flights on social media, a model could be generated that can be used to quickly evaluate what people are tweeting about. By have the sentiment of several thousand tweets. A model could be provided that is able to accurately guess the sentiment of future tweets related to flights. The purpose of the application is to be able allow not only companies but customers to see the sentiment of recent tweets relating to an airline.

Introduction

On Kaggle, there is a data set of Tweets regarding Airline Sentiment. These tweets are organized in a number of ways, such as positive, negative, or airline company. The type of project I am proposing is an application project. This project would use the 2015 Twitter Airline Sentiment dataset. With this dataset I could create a sentiment analysis application. The dataset also contains what the user tweeted about, whether it be the service, wait time, etc. By reading

the tweets and their associated sentiment a model could be trained. From there any tweet could be input into the application and the associated sentiment could be returned. Later on other features could be added, such as confidence and reason. The current version of the application uses logistic regression to provide the user with the sentiment of an inputted tweet. Later on, neural networks may provide a more accurate and consistent guess than logistic regression. This will make our user more confident is the outputted sentiment of our application.

Background

Looking through what other users have done with this dataset has been interesting. Most of the other users seem to only have analyzed the data and provided visualizations of the data. There are a couple standout individual submissions of this. Data visualization and analysis is nice but at the end of the data it goes over the head, or is too dry for many people. However, only a few users have actually attempted to use artificial intelligence on this dataset. This may be because most of the dataset is text and text can be difficult to work with. I want to create a simple application for the dataset that will provide basic feedback. The other's work will certainly be useful to use as a reference as I experiment and move forward due to my beginner understanding of artificial intelligence.

Methodology

Working with tweets more difficult that just using data points. To parse the tweets was a process. I first had to tokenize the tweets. This would turn each word into its individual token. Then all tokens that were just special characters or numbers were discarded. The tokens were

then lemmatized. After that the stopword were removed. Finally, the tokens were turned into a matrix and then vectorized. Once the tweets were parsed the method I ended up using was the logistic regression. I used a dummy classifier with a X train test split for the data to train the logistic regression. It provided fairly accurate results and was simple to implement. The method I wanted to use, if I had unlimited time, was the Multilayer Perceptron. Formatting the data for the multilayer perceptron was a difficult task that went over my head. Also the multilayer perceptron can take a long amount of time to work in my own environment due to how intensive it can be with computing power. Working with a such a large dataset of 17,000 tweets it takes a good amount of time just for logistic regression to run.

Experiments

Using logistic regression I was able to get an accuracy of around 0.77 or 77%. This is a fairly strong accuracy but it is not perfect. I ran experiments by asking my housemates for tweets. Many of them gave me absurd or silly things to test. Some of the tweets I tested were “I love this plane”, “I hate this plane”, “I love this plane. I hate this plane”, “I hate this plane. It's so sad. I’m going to cry.”.

Analysis

What I found from the experiments was interesting. Really obvious words, like love, like, sad or hate, seem to carry a lot of weight when deciding the sentiment. I also noticed that negative words seem to influence the tweet more than positive words. For example, “I love this plane. I hate this plane”, returned a negative sentiment rather than a neutral one. This may be

because sometimes negative tweets have positive words in them rather than positive tweets sometimes having negative words in them.

Conclusion

Analyzing the sentiment of tweets is a powerful tool. It can be used by companies or customers to see what people are saying about airlines. The results provided by logistic regression aren't the best but it is a solid proof of concept. Learning how to work with more complex datasets, such as tweets, was valuable. At first it was daunting but now I feel much more comfortable working with them. In the future, more powerful learning algorithms, such as neural networks, could be used on this dataset or similar ones.

.