Capstone – I

PROJECT PROPOSAL

Priya Theru

Springboard Data Science Career Track, April 2019 cohort

Problem Statement:

Classifying Twitter sentiments based on positive/negative/neutral labels using NLP

Impact :

Labels represent the sentiment of a text. These labels helps in analyzing the travellers' experience in using the US airlines - the actual feedback is from the text itself. The primary goal is to build a classifier that understands the text and assign an appropriate label to it. Background Each sentiment label is categorized as positive,negative and neutral. Therefore we will have a supervised, multi-class classifier with actual text as input. Background Each sentiment label is categorized as positive,negative and neutral. Therefore we will have a supervised, multi-class classifier with actual text as input. This piece of code is an exploration of Natural Language Processing (NLP). The goal of predicting the labels given a piece of text will deal with plenty of NLP topics such as NGrams, NamedEntityRecognition, Bag of Words. Finally, we arrive at a dataframe and we will be employing different machine learning algorithms on it to come up with the best model. The dataset contains the travellers' reviews on US Airlines in February 2015. There are a total of 14640 records of different airline reviews.

The data dictionary is as follows:

tweet_id - user's unique ID on the twitter , int64
airline_sentiment - sentiment labels, string
airline_sentiment_confidence - degree of the sentiment, float64
negativereason - reasons out the negative sentiment, string
negativereason_confidence - degree of the negative sentiment, float64
airline - Specified name of the airline, string
airline_sentiment_gold - sentiment labels of airline's gold members, string
name - Specified name of the user on twitter, string
negativereason_gold - reasons out the negative sentiments of airline's gold members, string retweet_count - twitter's retweet count,int64 text - user's reviews, string
tweet_coord - latitude and longitude coordinates, int64
tweet_created - time when the tweet was posted, string
tweet_location - place where the tweet was posted, string
user_timezone - time zone where the review was being posted, string.