Trump vs. Clinton: Big Data Analytics Insights

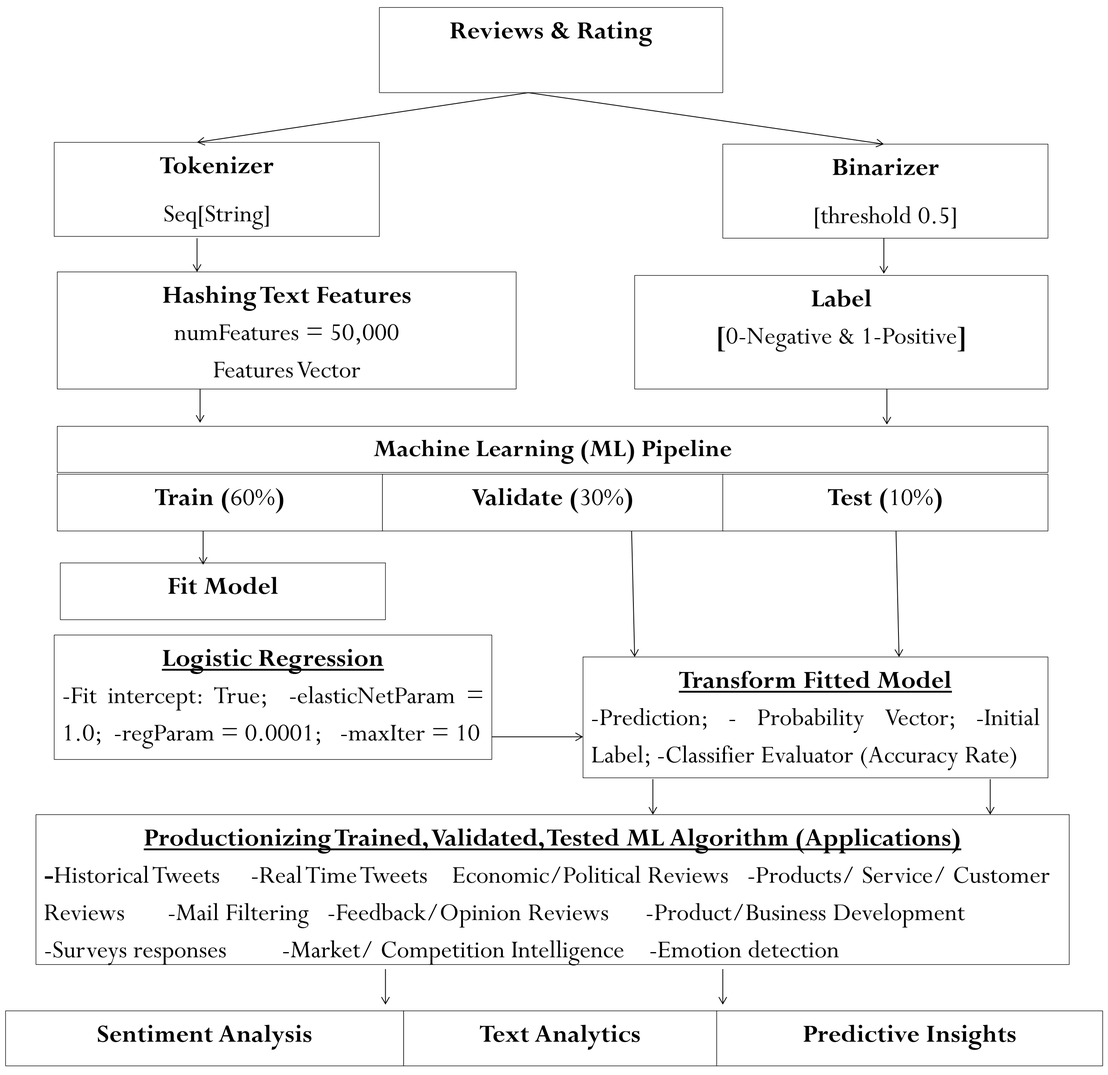
Akinwande A. Atanda
Department of Economics and Finance
https://nz.linkedin.com/in/akinwande-atanda https://github.com/aaa121
UNIVERSITY OF CANTERBURY

Introduction

The race among political parties candidates to win the 58th quadrennial United States presidential election of 2016 have generated several media attentions, reviews and comments by voters. The series of presidential primary elections across the 50 states made Donald Trump and Hillary Clinton became the presumptive presidential nominees of the Republican and Democratic Party respectively. The businessman, Trump and former secretary of State, Hillary both used Twitter as a key tool of their campaigns. Millions of daily tweets by potential voters and non-voters are about the candidates' economic and non-economic plans. Tweets by Donald Trump with the screen name @realDonaldTrump are being tagged as controversial, hilarious and funny. @HillaryClinton is known as being a conservative and non-controversial candidate.

The large pool of tweets in form of comments and reviews by voters and candidates provide potential diagnostic and predictive insights about the next U.S president by November 8, 2016. This type of analysis forms one of the applied areas of "Text Analytics" in the field of "Computational Programming". The application motivates this study to employ a Supervised Machine Learning (ML) task to analyse approximately 50.3million tweets collected between April 29 and June 23 2016 (8 weeks).

Algorithm Schema

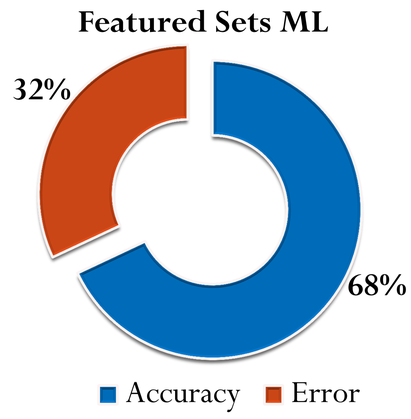


Method

The text data used for training the machine learning algorithm are sourced from the NLTK movies reviews and Amazon products reviews repository. The data are tokenized into sequence of strings and binarized using their ratings into positive and negative reviews as shown on the schema.

The big text data for featurization and productionalization are processed using a cloud based computing engine, Apache Spark through Databricks. The tweets are streamed using Spark Streaming API, filtered-[Extract-Transform-Load (ETL)]-with Spark SQL and fitted using binary based classifier ML algorithm. The applications used for the analytics are written in Scala, R, and Python.

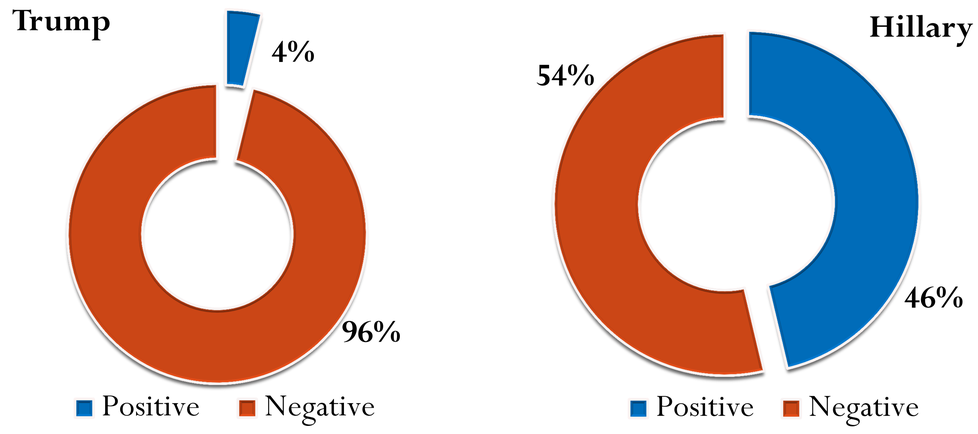
The features vector is randomly split into training (60%), validating (30%) and testing (10%) dataset. The training features set are fitted with a logistic regression model to predict the probability of the outcome being negative (0) or positive (1) label. The designed ML binary classification algorithm was trained, validated and tested for 36 hours without interruption. For the features set, the algorithm accuracy rate ranges between 0.582 and 0.773 while looping through the elastic net parameter from 0 to 1.0.



Sentiment Analysis

Sentiment analysis (SA) or opinion mining (OM) is conducted as a computational study of voters tweets about the presidential candidates. Two lists of keywords are used to filter the tweets for each candidate.

Filter: Keywords and Hashtags	
Donald Trump	"Trump2016", "#MakeAmericaGreatAgain", "Donald Trump", "#lovetrumpshate"
Hillary Clinton	"Hillary Clinton", "#neverhillary", "#hillaryclinton", "#crookedhillary"



The filtered tweets from the merged batches of streams are productionalized based on the trained, validated and tested classifier algorithm to detect which of the voters view is positive or negative opinion. Comparatively, 96.2% of tweets about Donald Trump are negative, while 53.7% for Hillary Clinton.

Economic Policy: Keywords/Text Analytics

Keywords search on Trump's Twitter page indicate less frequent use of economic policy related terms-["tax", "immigration", "education", "health care", "foreign policy", "economy", "jobs", "debts"]-in his tweets relatively to Hillary Clinton.

Predictive Insights

From each candidate Twitter profile as at 23rd June 2016, Donald Trump has the highest number of tweets (32,333) and followers (9.23 million) compared to Hillary Clinton (6,128; 7.05million). Despite Trump engineered his tweets to gain media and voters' attention, sentiment polarity about him has been mostly negative compared to Hillary. Relating the profile statistics with sentiment polarity results, it suggests that some anti-Trumps are following Trump on Twitter.

On the basis of economic policy keywords analytics and sentiment analysis, there is a high chance of Hillary Clinton emerging as the next US president. Using Multiclass algorithm evaluator with trained first 100 historical tweets, the accuracy rate stood at 86.4%.

Supported by: **Databricks Academic Partners Program & Amazon Web Services Educate**
databricks aws educate