PROJECT FLETCHER SUMMARY
TOPIC ANALYSIS OF TWEETS #MIDTERMS2018 #ELECTIONNIGHT

-Hiranya Krishna Kumar

DOMAIN

In Project Fletcher ,I collected 70,000 tweets with the hashtag #Midterms2018 and #ElectionNight from twitter using Tweepy, a python wrapper for twitter API using search API and stored the data in MongoDB using AWS. There were 12,931 unique tweets in total.The aim was to build a model to predict retweet_count given the topic of the tweet. Sentiment Analysis was done to understand the distribution polarity and subjectivity of the tweet

DATA

Field Name	Required	Definition
retweet_count	Yes	Max retweet count of a tweet
favorites_count	Yes	No. of likes
polarity	Yes	Polarity of tweet
subjectivity	yes	Subjectivity of tweet
Topic 0-9	Yes	Document- topic probability distribution
created_at	No	time the tweet was created
user_id	No	User id of tweet

PREPROCESSING

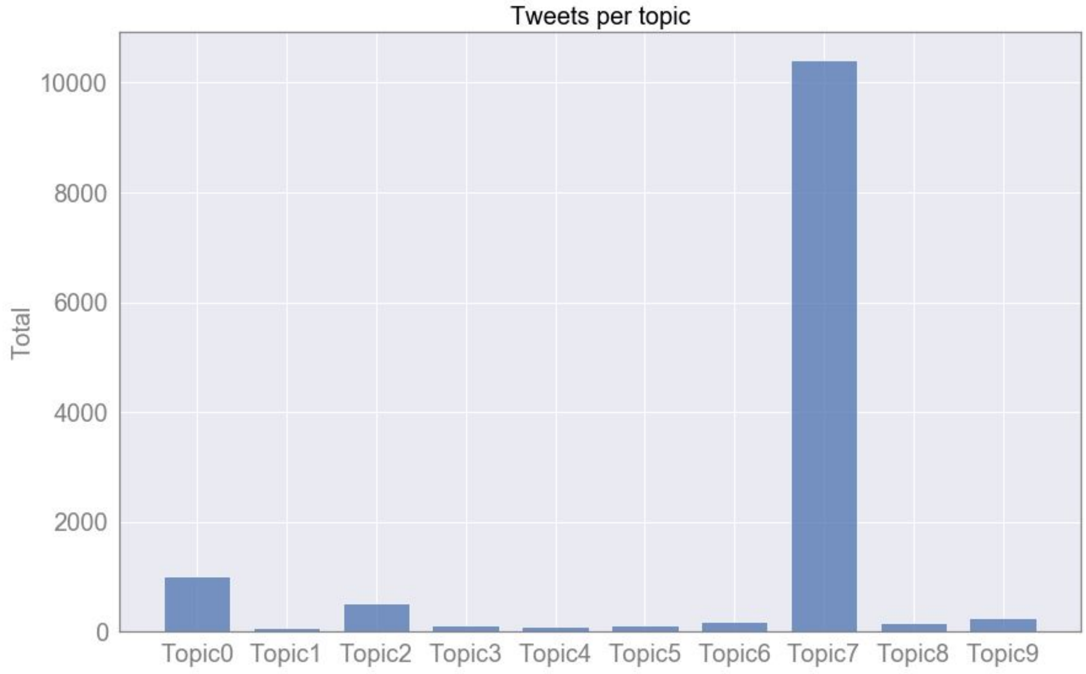
The tweets were first cleaned by removing punctuations and keeping only alphabet characters .All characters were converted to lowercase . Preprocessing tweets included removing stopwords and hashtags , stemming and tokenization.Sentiment analysis was done using Text Blob for all the tweets. There were 6670 positive ,4026 negative and 2022 neutral tweets.

TOPIC MODELING AND SUPERVISED LEARNING

Words in the tweet were mapped into a vector space using Term Frequency distribution of the words in the tweet with count-vectoriser . Unigrams , bigrams and trigrams of the words were used as a feature in the count vectoriser. Topic modeling was done using LDA , online learning method with a batch of 400 documents . Word topic distribution was set to 0.005 to get disparate topics and no of components was set to 10 topics. These hyperparameters were tuned so that the model perplexity was minimum and the words in all topics could be generalised into topic labels.
A gradient boosting regressor model was used to predict the retweet count using the doc-topic probabilities and sentiment of topics as features

RESULTS

The topics of the LDA model can be categorised broadly as : r historic wins, democrats flipping hose and texas senate elections,states voting to legalize marijuana, governor elections in georgia,recent trump firings,the most frequent topic was a general topic on elections that contained 3 other subtopics. All the topics were visualized using pyLDAvis. The distribution of documents for all topics is shown below

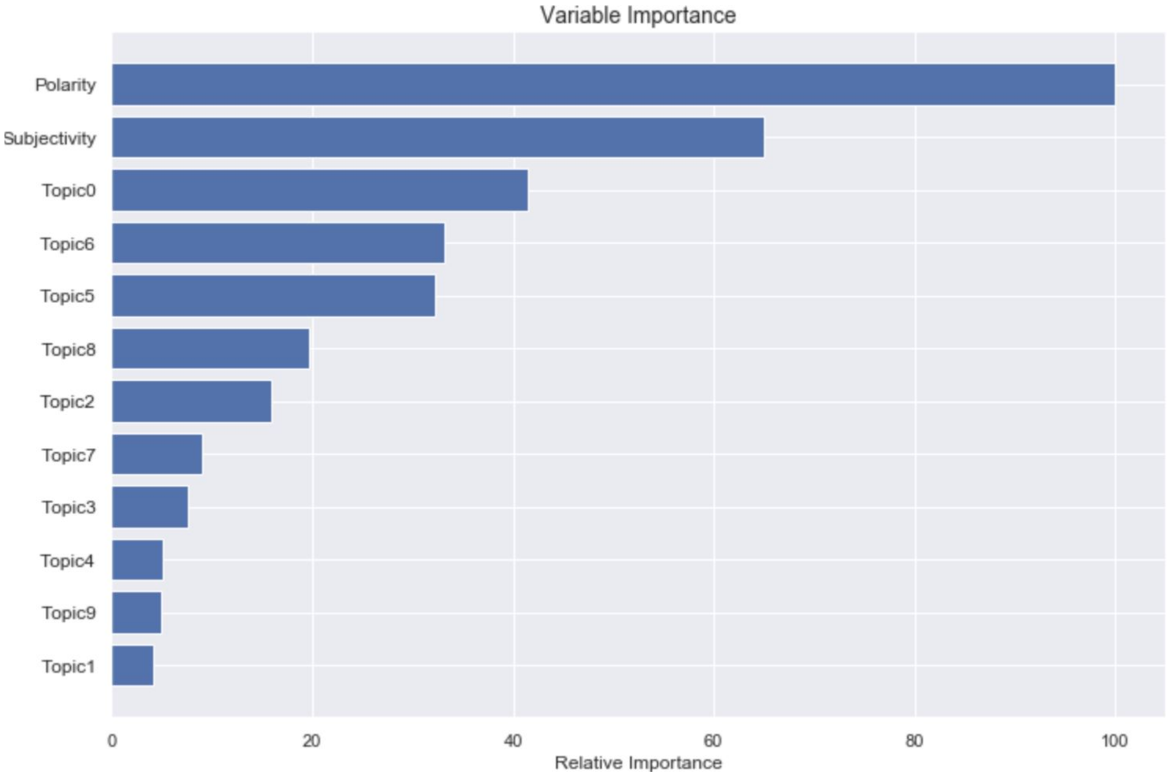


PROJECT FLETCHER SUMMARY
TOPIC ANALYSIS OF TWEETS #MIDTERMS2018 #ELECTIONNIGHT

-Hiranya Krishna Kumar

CONCLUSIONS:

Using gradient boosting regression to predict retweet counts, the most important topics were Topic 0: historic wins, topic 5: states legalising marijuana and topic 6: recent Trump firings.



FUTURE WORK

1. Use a different pipeline like word2vec and k-means clustering to identify topics and most frequent words in tweets
2. Gather more tweets and use guided LDA to get subtopics within a topic
3. Tune hyperparameters like doc_topic_prior,evaluate_topics frequency,word_topic_prior
4. Build a flask app to predict a topic for a tweet on midterm elections