Twitter Sentiment Analysis

Shahbaz Masood
Springboard Capstone Project 2

Problem & Motivation

- Sentiment analysis is a multidisciplinary necessity.
  - Online Commerce
  - Voice of the Market (VOM)
  - Voice of the Customer (VOC)
  - Brand Reputation Management
  - Government

- Extract sentiment from 100,000 tweets.
- Dataset acquired online

Data

- Publicly available.
- Data
  - 100,000 tweets
  - Each marked with 0,1 i.e negative positive sentiment

sentiment	tweet
1	@awaketoday awww, thanks
1	@2NiteBoy damn it xD hurry up
1	@celiabb 8 but the next one up would be better
1	@cleeobea drink lots of water- they always help my headaches not seem so intense! and then go to bed!
1	@biggsjm Is it a family tradition?
0	@CrazyBallerina can't. my sal wont be in until next week
0	#followfriday @SusieGennoe - she's only got 35 followers and that makes me so sad
1	@612brisbane try following my public account instead of @rocketpilot
0	@BarackObama don't act like you didn't smoke in your hayday, buster!! Don't take my flavors away
0	I got scolded for not waiting and spending MORE to find my perfect storage solution... saddies Guess I should have *hangs head*

Data Preprocessing

The tweets were processed according to the following and in the same order.

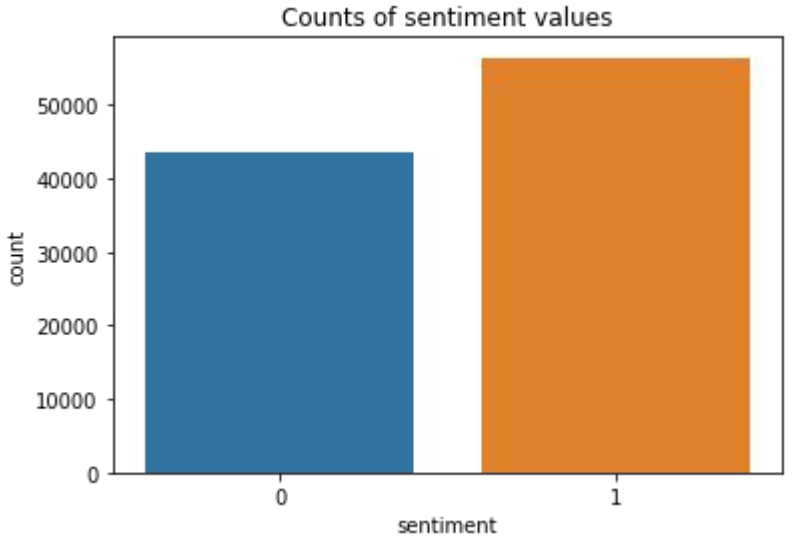
1. Tags removal (i.e '@')
2. Lowercasing
3. Numbers removal
4. HTTP links removal
5. Emojis were processed. Emojis were labelled EMO_POS or EMO_NEG
6. Punctuation removal
7. Removed extra white spaces
8. Words which did not consist of alphabets were removed
9. Stop words were removed (These words add no value to the sentiment of the tweet e.g The, He etc)
10. Character repetitions were removed e.g funnny was changed to funny
11. Words were lemmatized to bring to their basic form e.g adventurous changed to adventure

EDA

Distribution of Sentiment

Roughly 56% tweets are marked positive and 44% as negative

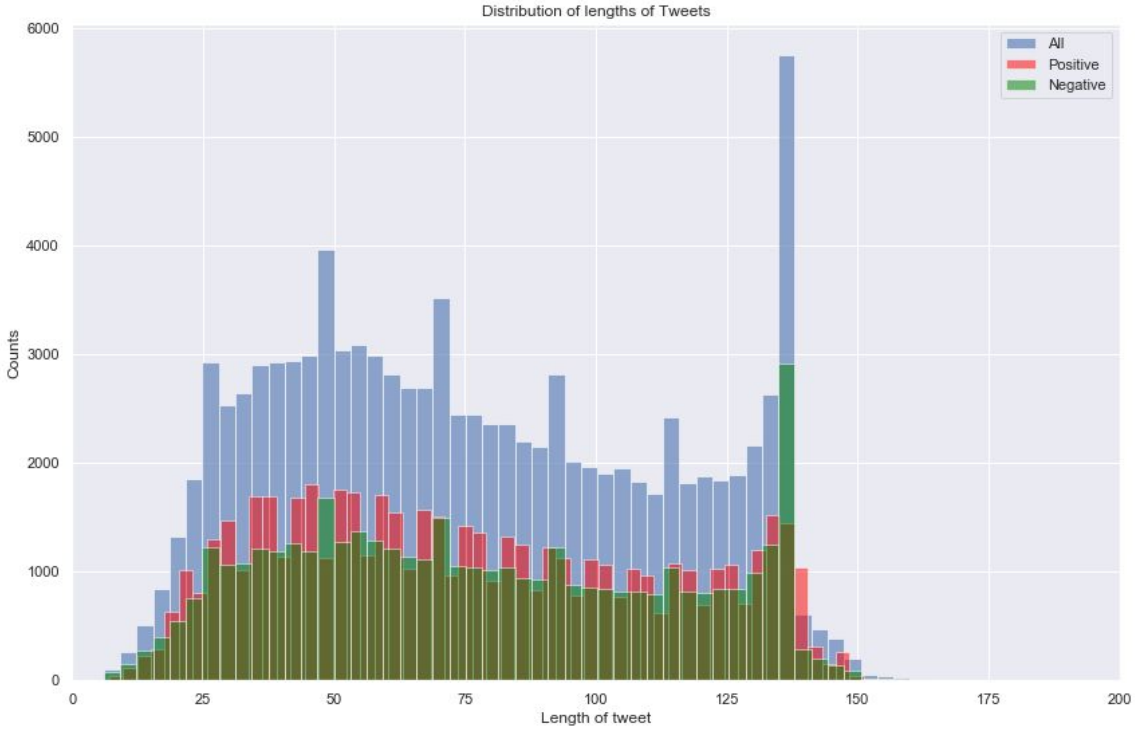
Low class imbalance



EDA

Distribution of length of tweets

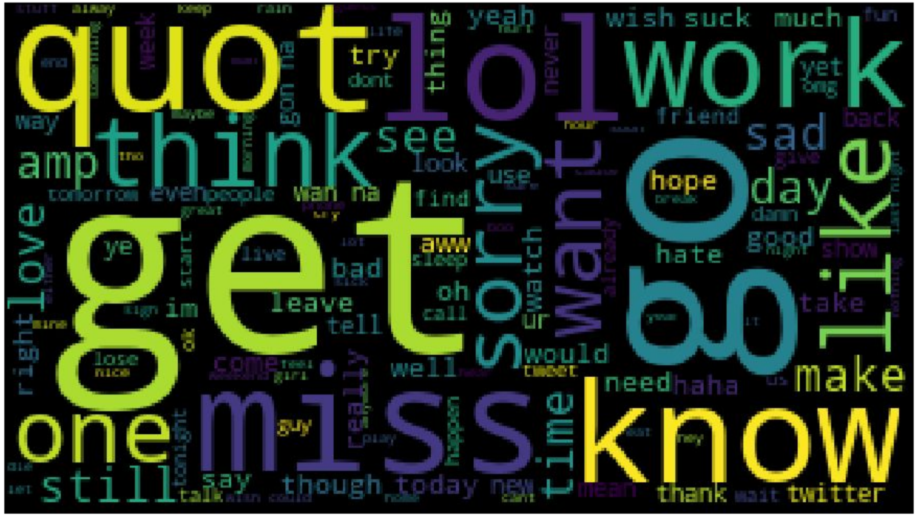
Equal for both sentiments

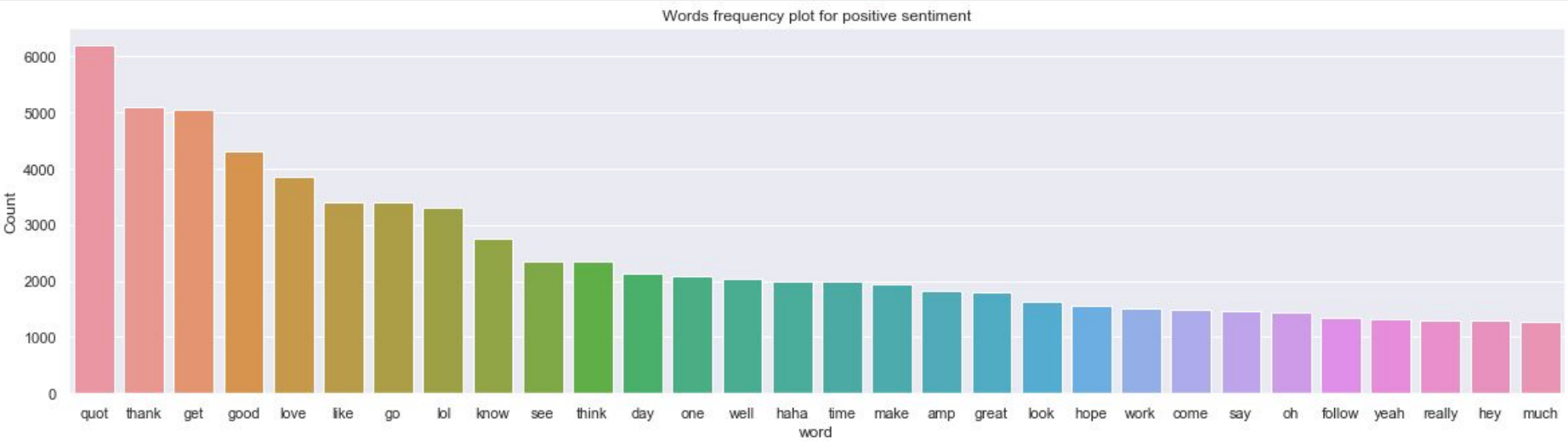


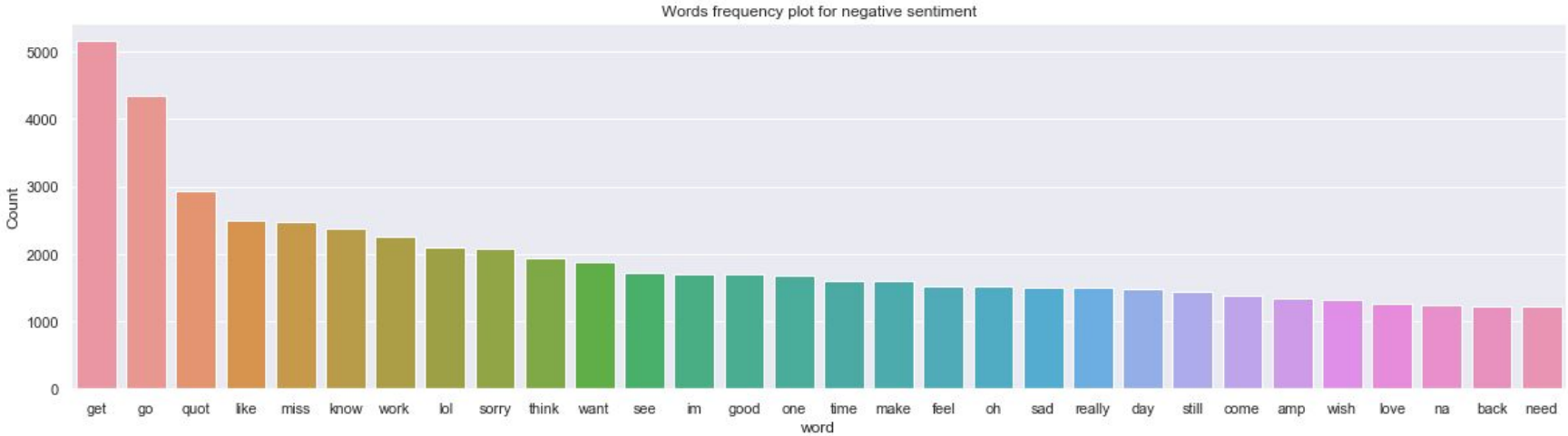
Wordclouds and Countplots

Positive tweets wordcloud

Negative tweets wordcloud




Words frequency plot for positive sentiment
Count
0
1000
2000
3000
4000
5000
6000
quot thank get good love like go lol know see think day one well haha time make amp great look hope work come say oh follow yeah really hey much
word


Words frequency plot for negative sentiment
Count
0
1000
2000
3000
4000
5000
get go quot like miss know work lol sorry think want see im good one time make feel oh sad really day still come amp wish love na back need
word

Initial problems with the data

Upon visual inspection it was noticed that some tweets are:

1. Marked incorrectly e.g. *smoke a little for 420 and chill out!*
2. Had no sentiment but a sentiment is assigned e.g. *called at the Farmers Market in the rain this morning, didn't buy coffee, sorry, I'm a tea drinker really!*
3. Some tweets had multiple sentiments i.e. one sentiment could not be assigned in entirety. E.g. *'i wanted to win but it's allll good lol'*

Modeling Overview

- 3 Text Feature Extraction methods
    - Bag of words
    - TF-IDF
    - Word Embeddings
- 6 ML models
    - Logistic Regression
    - SVM
    - XGBoost
    - NaiveBayes
    - DecisionTree
    - RandomForest

Data split scheme

	All data	Training data	Test data
Percentage	100%	80%	20%
Tweets	100,000	80,000	20,000

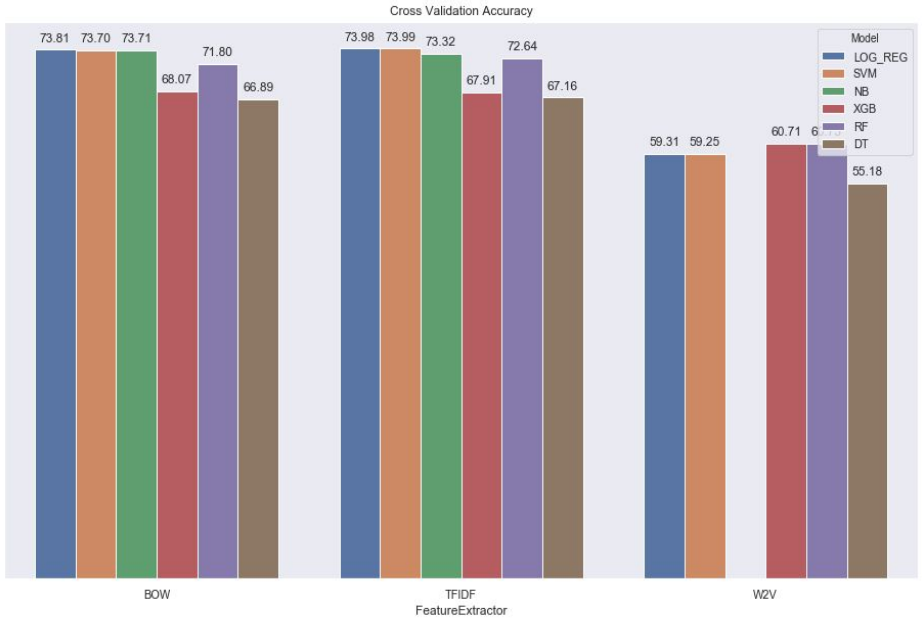
- 3-Fold cross validation
- Hyperparameter tuning
- Homogeneity maintained to ensure comparable results

Results

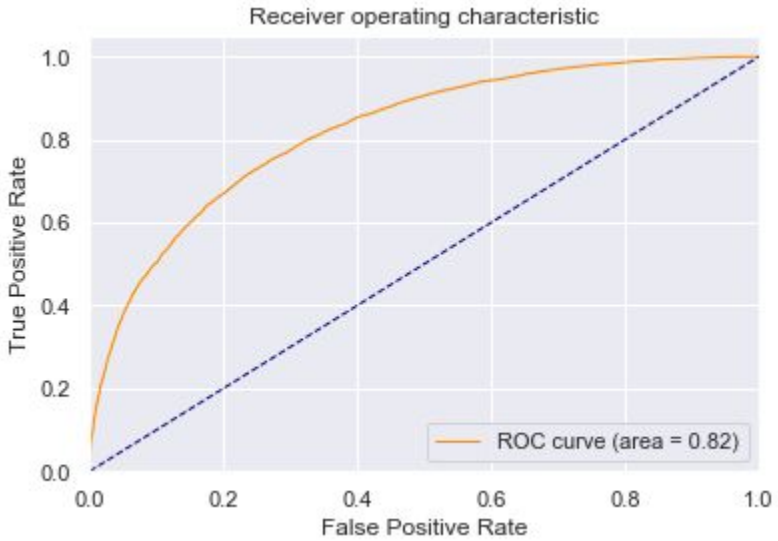
TFIDF - SVM wins all metrics

	FeatureExtractor	Model	TrainAccMean	TrainAccStd	CVAccMean	CVAccStd
0	TFIDF	SVM	0.767643	0.000807	0.739871	0.000922
1	TFIDF	LOG_REG	0.768249	0.000805	0.739808	0.000542
2	BOW	LOG_REG	0.769830	0.001403	0.738096	0.000265
3	BOW	NB	0.758942	0.000581	0.737095	0.001377
4	BOW	SVM	0.769580	0.000605	0.736983	0.000849
5	TFIDF	NB	0.758217	0.000717	0.733157	0.000327
6	TFIDF	RF	0.976710	0.000521	0.726419	0.002346
7	BOW	RF	0.976885	0.000493	0.717993	0.001447
8	BOW	XGB	0.685565	0.002639	0.680652	0.004457
9	TFIDF	XGB	0.686258	0.003049	0.679139	0.004392
10	TFIDF	DT	0.976722	0.000514	0.671601	0.002833
11	BOW	DT	0.976916	0.000491	0.668888	0.001787
12	W2V	RF	0.988205	0.000353	0.607343	0.004190
13	W2V	XGB	0.626758	0.000899	0.607056	0.004873
14	W2V	LOG_REG	0.595142	0.001879	0.593117	0.005842
15	W2V	SVM	0.595417	0.001232	0.592504	0.005994
16	W2V	DT	0.988205	0.000353	0.551787	0.003365

CV Accuracies



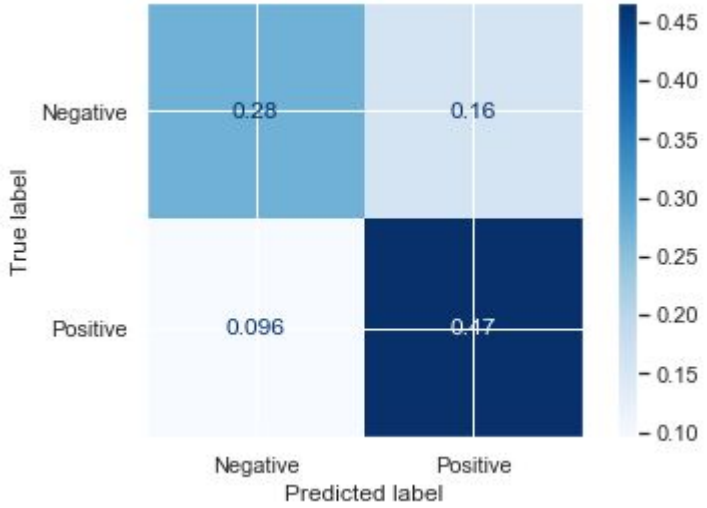
ROC of best model



Confusion Matrix

- We can see that 16% of the error is because of False Negatives while 9.6% is due to False Positives.

- The lower f1-score for Negative sentiments indicates that it is relatively harder to predict a negative sentiment correctly



	Negative	Positive	micro avg	macro avg	weighted avg
f1-score	0.598614	0.787332	0.721972	0.692973	0.704759
precision	0.812623	0.690924	0.721972	0.751773	0.744172
recall	0.473829	0.915007	0.721972	0.694418	0.721972
support	8750.000000	11248.000000	19998.000000	19998.000000	19998.000000

Further Improvements

1. Use more features to train the models for TFIDF.
2. Use a sophisticated model like deep learning where the models are trained with the context of a window of words.
3. BOW and TFIDF can be used with 2-grams or 3-grams
4. Training dataset can be refined to be more accurate.
5. Exclamation marks could be handled specially during the preprocessing of the tweets.