Yelp Spam/Review Detection

Presented By
Hansraj Pabbati (SJSU ID : 012540541)
Rishika Machina (SJSU ID: 012525227)

Introduction

- Online reviews play a very important role for decision-making in today's e-commerce.
- Untruthful review is a fake review or opinion spam
- Positive reviews of a target object can attract more customers and increase sales; negative reviews of a target object can result in lower demand and lower sales.

Filtering

Supervised

Unmonitored

Characteristic

Linguistic

Behavioral

	date	reviewID	reviewerID	reviewContent	rating	usefulCount	coolCount	funnyCount	hotelID	Target
0	6/8/2011	MyNjnxzZVTPq	IFTr6_6NI4CgCVavIL9k5g	Let me begin by saying that there are two kind...	5	18	11	28	tQfLGoolUMu2J0igcWcoZg	0
1	8/30/2011	BdD7fsPqHQL73hwENEDT-Q	c_-hF15XgNhlyy_TqzmdaA	The only place inside the Loop that you can st...	3	0	3	4	tQfLGoolUMu2J0igcWcoZg	0
2	6/26/2009	BfhqiyfC	CiwZ6S5ZizAFL5gypf8tLA	I have walked by the Tokyo Hotel countless tim...	5	12	14	23	tQfLGoolUMu2J0igcWcoZg	0

Dataset

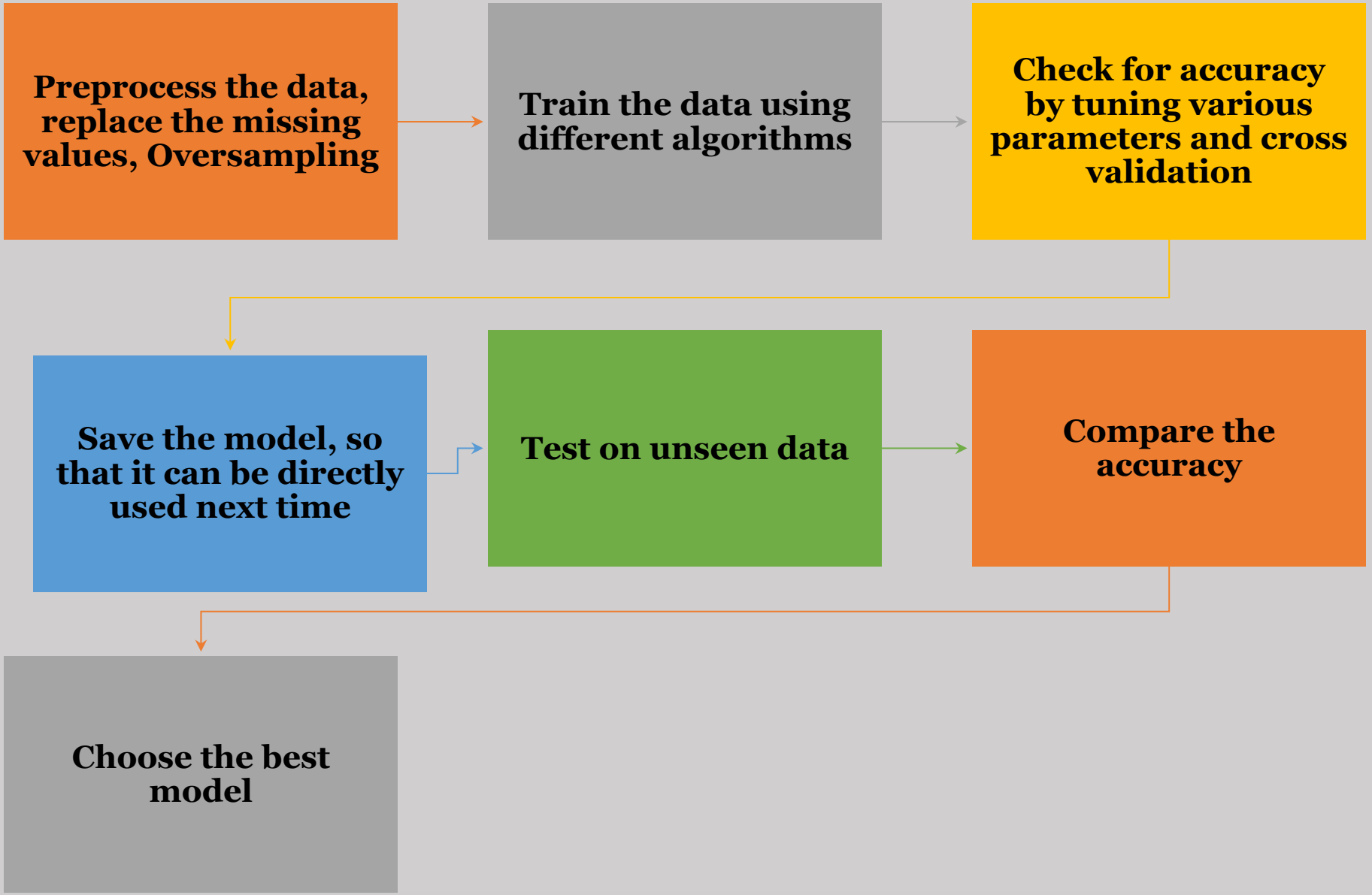
reviewLength	sentimentPolarity	sentimentSubjectivity
1249.0	0.248854	0.491878
593.0	0.075368	0.485091
90.0	-0.025000	0.233333
272.0	-0.021429	0.478571
151.0	0.400000	0.566667

- Preprocessing: Tokenize reviews (200000-dimension feature vector)
- Fabricated features from Review Content
- Used Over sampling to balance the classes using XGBoost

Dataset

System Design and Implementation

- Structural Features - Review Length, Average and standard deviation of review length
- Semantic Features – Sentiment Polarity
- N-Gram - Unigram, Bigram and Trigram


Preprocess the data, replace the missing values, Oversampling
Train the data using different algorithms
Check for accuracy by tuning various parameters and cross validation
Save the model, so that it can be directly used next time
Test on unseen data
Compare the accuracy
Choose the best model

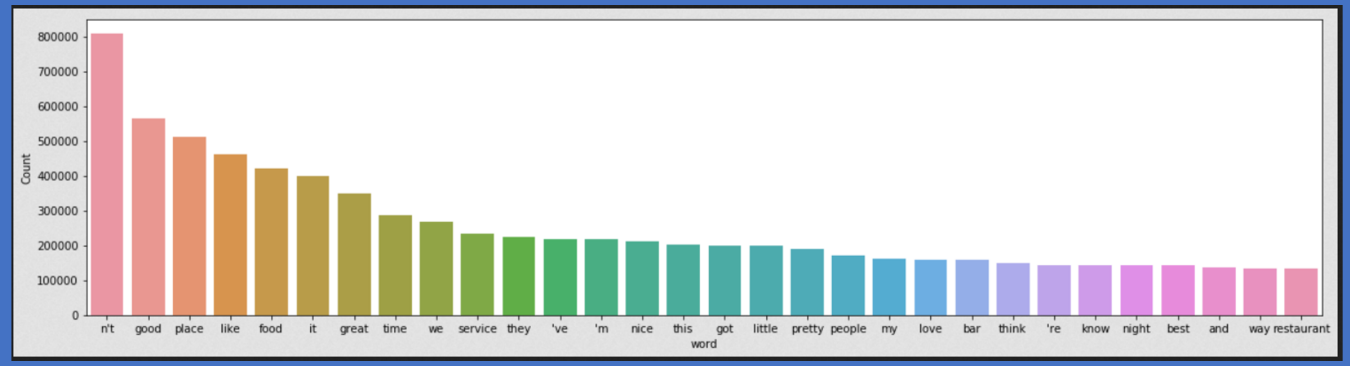
Data Preprocessing and Feature Engineering

- Train – Test Split 80-20 percent with 5 fold cross validation
- Tokenize using Spacy and Keras
- Removal of punctuations , convert to lowercase, Regex
- Removal of Stopwords
- Tokenization and Lemmatization
- Vectorizer and TF-IDF
- Word Frequency and Word cloud
- Sentiment Polarity and Sentiment Subjectivity

Word Cloud

Heat Map to find Correlation

Most Frequent Words



Multinomial Naïve Bayes

- Trained on Review Content initially
- Gave decent results, so to improve it further used Grid Search
- Performed Grid Search with following parameters-
  'vect__ngram_range': [(1, 1), (1, 2)],
  'tfidf__use_idf': (True, False),
  'clf__alpha': (1e-2, 1e-3)}
- *Best Parameters obtained :*
  ngram_range':, (1, 2),
- *Tfidf__use_idf : True, clf_alpha : .0001.*
- 5 – fold cross validation

Random Forest

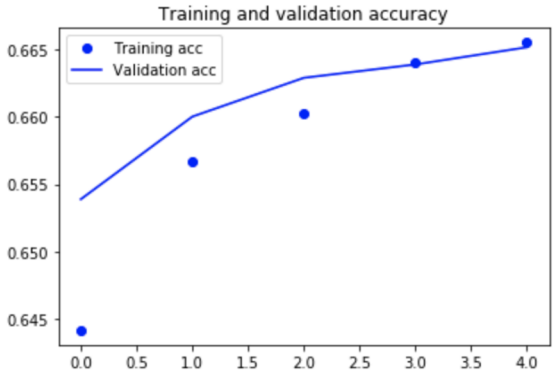
- 501 estimators
- balanced class weights - Imbalanced dataset
- Bootstrap aggregation

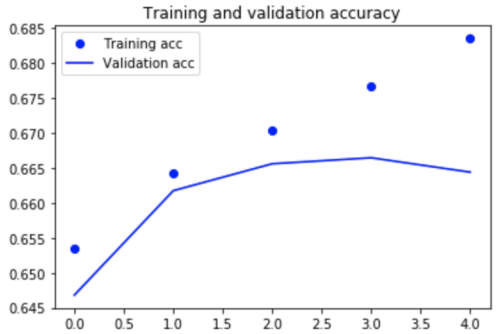
- CART - Gini Index
- ID3 - Information gain

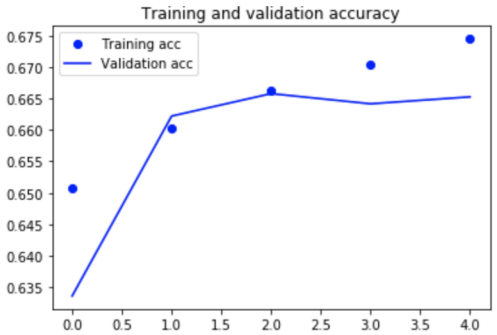
Neural Networks

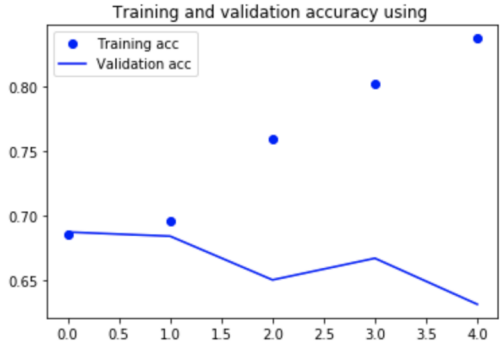
Neural Network with LSTM	Neural Network with extra 1D CNN on top of LSTM layer	Neural network with Bidirectional recurrent layers	Neural network using RNN with GRU
Embedding layer, LSTM layer, and output of LSTM Layer was fed into hidden fully connected layer	Combination of recurrent neural network with LSTM and Convolutional layer	Frequently used for NLP for richer data representations	Trained with padded sequence
10,000 most common words were used as features	Dropout layer was added after embedding layer and convolutional layer	Bidirectional recurrent layer after embedding layer	Need to change weight matrices so recurrent unit provides output for input sequence
20 percent dropout, 0.2 Validation split	20 percent dropout, 0.2 Validation split	Separate instance of this layer for chronological processing and reversed order processing	5 percent validation split

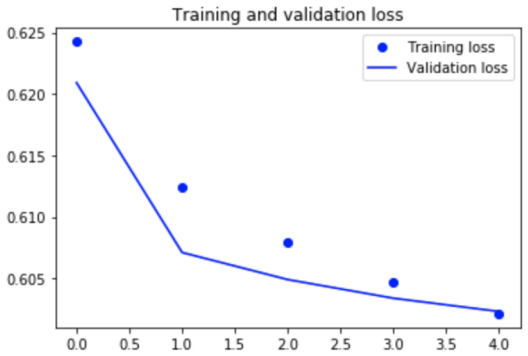
Neural Network with LSTM	Neural Network with extra 1D CNN on top of LSTM layer	Neural network with Bidirectional recurrent layers	Neural network using RNN

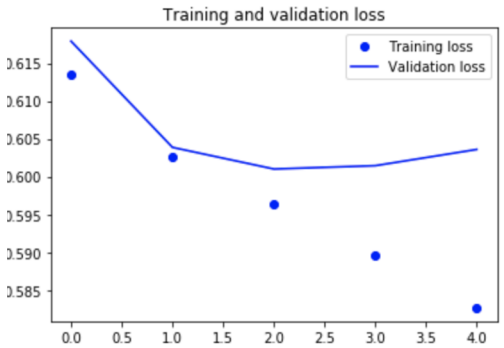


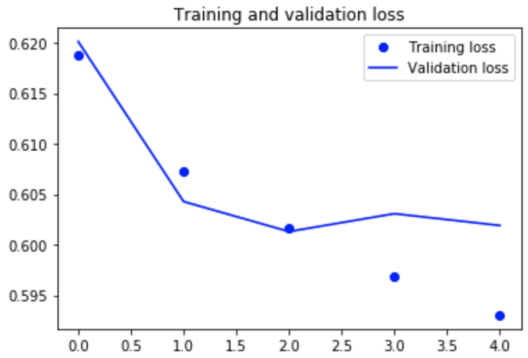


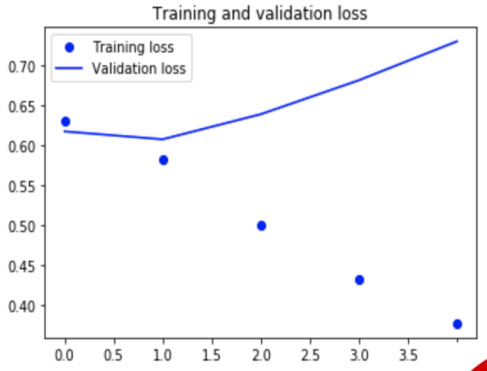








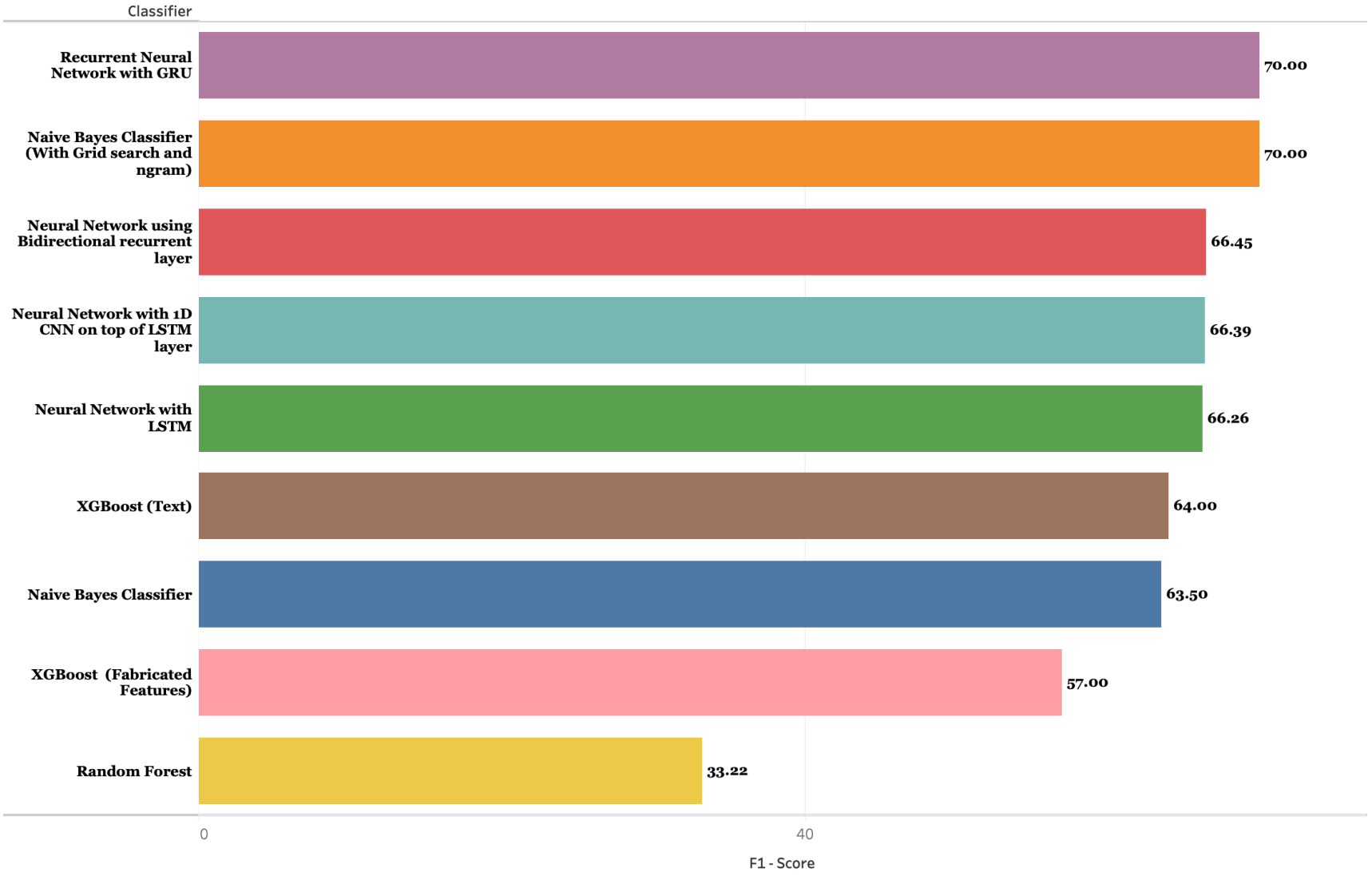




Evaluation of Results

	Classifier	F1 Score
1	Naive Bayes Classifier	63.52 %
2	Naive Bayes Classifier (With Grid search and ngram)	70.00 %
3	Neural Network with LSTM	66.26 %
4	Neural Network with 1D CNN on top of LSTM layer	66.39 %
5	Neural Network using Bidirectional recurrent layer	66.45 %
6	Recurrent Neural Network with GRU	70%
7	XGBoost	64% (Text), 57%(Other)
8	Random Forest	33%

F1 Score Evaluation



Conclusion

- No major improvement with reviewer centered fabricated features.
- Using simple approach works best.
- Multinomial Naïve Bayes gave the best accuracy of 70 percent with Grid search.
- RNN with GRU also gave 70 percent accuracy

Thank you