

Sentiment Analysis of Reviews

Creative Component Project Report

By:

Pravalika Ravikumar Chitla

Master of Science in Information Systems

Submitted in fulfillment for the requirements for the degree of Master of Science in
Information Systems

Major Professor: Dr. Anthony Townsend

Ivy College of Business

Iowa State University

Ames, Iowa

2021

Table of Contents

Acknowledgement.....	6
Abstract.....	7
1. Introduction.....	8
2. Literature Review.....	9
3. Project Methodology	10
3.1 Data Source.....	10
3.2 Exploratory Data Analysis.....	10
3.3 Data Preprocessing.....	13
3.3.1 Removing reviews written language other than English.	14
3.3.2 Feature Selection - Selecting important columns.....	15
3.3.3 Removing Punctuations, non-alphanumeric characters, lowercase conversion, tokenization.....	16
3.3.4 Removing stop-words.....	18
3.3.5 Generating embeddings for each review.....	19
3.3.6 Standardizing review-length by truncating long reviews and padding short reviews.....	20
3.4 Selection of Machine Learning Models.....	21
3.4.1 Naïve Bayes classifier.....	21
3.4.2 Bernoulli Classifier.....	21
3.4.3 Logistic Regression.....	21
3.4.4 Stochastic Gradient Descent (SGD)	22
3.4.5 Support Vector Machine (SVM)	22
3.4.6 Long Short Term Memory (LSTM- a deep learning approach)	22
3.4.7 Designing LSTM model.....	22
3.4.7.1 Models and Hyperparameters Tuning.....	23
3.4.7.1.1 LSTM Model.....	23
3.4.7.1.2 Hyperparameters Tuning.....	23
4. Evaluation of Models	24

4.1 Evaluation Matrix.....	24
4.2 Results of different classification model.....	25
4.2.1 Naïve Bayes Classifier.....	25
4.2.2 Bernoulli Classifier.....	26
4.2.3 Logistic Regression.....	27
4.2.4 Stochastic Gradient Descent (SGD)	28
4.2.5 Support Vector Machine (SVM)	29
4.2.6 Long Short Term Memory (LSTM- a deep learning approach)	30
5. Comparing Evaluation Scores	31
6. Conclusion.....	32
References	32

List of Figures

Fig: 1 Graph for Number of reviews per year.....	11
Fig: 2 Graph for Number of reviews per month for the year 2018.....	11
Fig: 3 Graph for Distribution of Customer Rating.....	12
Fig: 4 Graph for Average Monthly Customer Rating for the year 2018.....	13
Fig: 5 Word Cloud for highest frequency words.....	18
Fig: 6 Tokenized words with word index.....	19
Fig: 7 GloVe embedding matrix.....	19
Fig: 8 Histogram of Review Lengths.....	20
Fig: 9 LSTM Model.....	23
Fig: 10 Graph for sentiment analysis using Naïve Bayes Classifier.....	25
Fig: 11 Graph for sentiment analysis using Bernoulli Classifier.....	26
Fig: 12 Graph for sentiment analysis using Logistic Regression.....	27
Fig: 13 Graph for sentiment analysis using Stochastic Gradient Descent.....	28
Fig: 14 Graph for sentiment analysis using Support Vector Machine.....	29
Fig: 15 Graph for sentiment analysis using LSTM.....	30
Fig: 16 Comparison of Evaluation Scores.....	32

List of Tables

Table:1 List of Reviews in other languages.....	14
Table:2 Dataset after feature selection.....	15
Table:3 Dataset before cleaning.....	16
Table:4 Dataset after feature selection.....	17
Table:5 Reviews converted into lowercase.....	17
Table:6 Summarized view of hyperparameters.....	24
Table:7 Table for sentiment analysis using Naïve Bayes Classifier.....	25
Table:8 Table for sentiment analysis using Bernoulli Classifier.....	26
Table:9 Table for sentiment analysis using Logistic Regression.....	27
Table:10 Table for sentiment analysis using Stochastic Gradient Descent.....	28
Table: 11 Table for sentiment analysis using Support Vector Machine.....	29
Table: 12 Table for sentiment analysis using LSTM.....	30
Table: 13 Table for analysis of evaluation values of all model.....	31

Acknowledgement

Foremost, I would like to express my sincere gratefulness to Dr. Anthony Townsend, my major professor for his constant assistance and guidance throughout my graduate studies and related research. This project would not have been achievable without his precious support and guidance.

Additionally, I would like to thank the graduate college faculty and staff. And all the professors for their constant support and guidance during my graduate studies.

Also, I would like to extend my thankfulness to my family, friends, and peers for their constant emotional support, timely encouragement, and boundless patience.

Abstract

The process of understanding a given text's emotion that one can use to analyze the product/service reviews is called sentiment analysis. Furthermore, classify them as either Positive, Neutral, or Negative sentiments. Learning the human masses' sentiments towards distinct entities and products facilitates better contextual advertisements, recommendation systems, and market trends analysis. Sentiment Analytics extract sense from human speech and have the power to suggest business insights into large amounts of data such as the opinions of customers, insights that inspire business decisions. Sentiment analysis systems allow companies to make sense of the sea of unstructured text by automating business processes, getting actionable insights, and saving hours of manual data processing by making businesses more efficient. I have chosen the Yelp dataset focusing on reviews of restaurants in Illinois to determine whether they are positive or negative sentiments and develop a model for their classification. This research project concentrates on predicting whether customers have a positive or negative attitude towards a restaurant. I have built a Machine learning model LSTM and used other classifier models such as Naïve Bayes, Bernoulli's, Logistic Regression, Stochastic Gradient Descent, and Support Vector Machine to train the models. I have used evaluation matrices such as prediction, accuracy, recall, and F_1 score to evaluate the classifier's performance on the test sample. Based on the evaluation matrix results, the goal is to find trends in how positive and negative reviews are written and come up with a model for their classification. The results show that SVM is the best model for classifying the reviews to either positive or negative as it gives the highest precision compared to the other classification models.

1. Introduction

Traditionally we used to learn about different products, restaurants, services, movies, hotels, etc. Through word of mouth, i.e., we get to know about these things based on our family members' experience or through our friend's circle. This process is very restricted; people get to know something either when they experienced it personally or discovered it from someone. This information can be broadcasted only across our friends or friends of friends where the reach is restricted. This way, there could be numerous things that not many people know, and that could be because very few people know about them, or people might have never experienced them. In the past few years, technology has advanced so much that we have started using it in every aspect of the business, such as social media, online shopping, traveling, etc. We have used technology to write our feedback and keep track of these users' reviews to understand how their experience was either good, bad, or neutral through their choice of words using sentiment analysis.

When you express your emotions and sentiments to people on a larger scale with the web's help, it becomes a powerful tool to communicate and convey your information. Sentiment analysis is the area of discipline that examines people's feelings, thoughts, emotions, evaluations, sentiments, and opinions towards entities such as services, products, organizations, individuals, issues, restaurants, events, movies, and topics. In plain terms, it helps in tracking the mood of the public. It uses natural language processing and data mining methods to the problem of gathering opinions from the text. Data creation is to reach 48 zettabytes by 2022. Every year nearly 2.5 quintillion byte of data is created, and it is estimated to double every year. One may consider this outbreak of data is significant. However, massive information can result in data overload. It isn't easy to extract the gold from the dross. In the data mining field of study, extracting and identifying patterns in massive data sets involves machine learning, statistics, and database systems to identify the text content. In today's world, the Internet's use has increased, and people have started to exchange opinions with the web's help, which has become the driving force for Sentiment Analysis. The web is a massive depository of structured and unstructured data. The analysis of this data to extract hidden public opinion and sentiment is a challenging task.

Sentiment analysis/Text analysis is the process of learning the emotion or opinion of a given text that one can use to analyze the product/service reviews. And classify them as either Positive, Neutral, or Negative sentiments. Sentiment Analytics/Text analytics extract meaning from human language and have the power to offer businesses insights into large amounts of data such as the opinions of customers, insights that can be used to drive business decisions. I want to extract customers' opinions from their reviews, help businesses understand their customers' needs, and further help businesses to draw insights to develop their business. So, the goal is to find trends in how positive and negative reviews are written and come up with a model for their classification. Sentiment analysis systems allow companies to make sense of the sea of unstructured text by automating business processes, getting actionable insights, and saving hours of manual data processing by making businesses more efficient. Yelp dataset is an excellent source of positive and negative feedback from customers that companies can leverage to enhance their services. On this vast dataset, the aim is to find trends in how positive and negative reviews are written and come up with a model for their classification. However, due to vast amounts of reviews on the

Yelp dataset, it can be time-consuming to read all reviews or compare other businesses' feedback. Hence, I have chosen the data reviews of restaurants in Illinois to determine whether they are positive or negative sentiments. This project further covers how data cleansing is done. It also explains how the sentences are tokenized and how to generate embeddings for each review. Further, it covers standardizing review length for long reviews. Additionally, it describes how the machine learning model LSTM is built and how different libraries are used for machine learning models. The various machine learning models used are Naïve Bayes, Bernoulli's, Logistic Regression, Stochastic Gradient Descent, and Support Vector Machine to train the models and then test the performance of the models on the test sample. Finally, it shows how comparison is performed between the models using the evaluation matrices like precision, recall, f₁, and accuracy to determine the best model for classifying positive and negative reviews.

2. Literature Review

A few techniques used for Sentiment Analysis are as follows:

For sentiment classification, Mori Rimon [3] used the keyword-based approach. He identified keywords that are essentially adjectives indicating sentiments. These indicators can be developed manually or can originate from Wordnet.

Alec co [4] used several machine learning algorithms. Different machine learning algorithms such as Naïve Bayes', Maximum entropy, and Support vector machine.

Janice M. Weibe [5] presented document and sentence level classification. He pulled review data from different product targets such as automobiles, banks, movies, and travel. He categorized the words into positive and negative classes. Further, he calculated the overall positive or negative score for the text. If the quantity of positive words is higher than negative, then the document is considered positive, otherwise negative.

Jalaj S. Modha, Sandip J. Modha, and Gayatri S. Pandi [6] aused different approaches in handling both subjective and objective unstructured data.

Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1 (2012): 1-167.

Pang and Lee [8] have explicitly worked on sentiment analysis and mining tweets from Twitter.

DENG et al. (2011) used sentiment analysis for Stock Price Prediction in [9].

K. Ming Leung [10] describes the Bayes rule.

Yaying Qiu et al [11] constructed extend Bayes model by assigning weights to important features.

In this paper, I have worked on identifying positive and negative reviews. Further, I have used different machine learning models and compared their performances to identify the best classification model for reviews.

3. Project Methodology

3.1 Data Source

The data used in this project is downloaded from Yelp. Yelp dataset consists of nearly 5.2 Million User reviews on numerous businesses from 11 metropolitan areas spanning across four countries. The primary focus is on the reviews of restaurants in Illinois along with star ratings and the city information.

Consider some examples of the Yelp dataset; these reviews are plain text in natural language and have some slang and idiomatic words. Here it consists of both negative and positive reviews that are distinct from each other. The classification models can be trained to learn the sentiment of the user.

– “Food is very bland — not authentic at all. meant to cater to the customers who have never eaten Vietnamese food before. Definitely will not be returning!” — 1 Stars.

– “A-M-A-ZING!!!!!!!!!!!!!! LOVE THIS PLACE!!!!!!!!!! Everything on the menu looked so good!! Garlic chicken was the BOMB!!!!!! MUST EAT AT THIS PLACE!!! I recommend ordering ahead before you go! Gets very busy!!!” — 5 Stars.

– “burgers are very big portions here. definitely order the onion ring tower to share...Milkshakes are tasty! My personal favorite — the vanilla one.” — 3 Stars.

– “Best Chinese on campus. Low prices for a ton of food. I get general taos chicken with fried rice, an egg roll, and a drink for under \$7.” — 5 Stars.

These reviews consist of slang, idiomatic words, punctuations, etc., that need to be cleaned and curated to use for our analysis.

3.2 Exploratory Data Analysis

Visualization is a great technique to do exploratory data analysis. So, I reviewed data from 2010 to 2019. Customers have started taking reviews for restaurants seriously, and the quality of reviews has significantly improved over time. So, it is essential to use the latest review data of the year 2018 to keep the model tuned for newer reviews for which the results are to be classified as “positive” and “negative” classes.

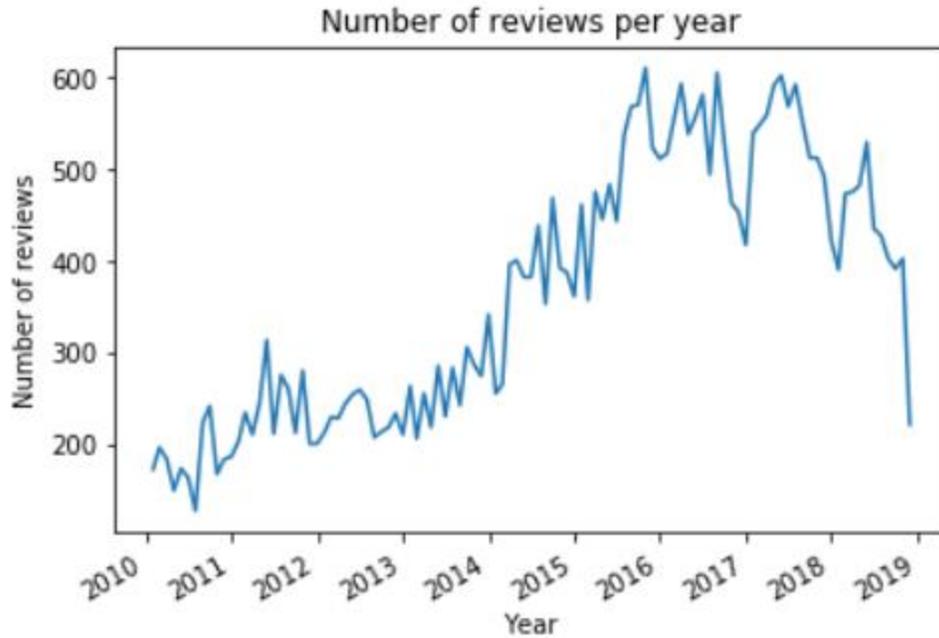


Fig:1 Graph for Number of reviews per year

I analyzed the number of reviews received every month in the year 2018, and as shown in the line graph, we see that in June, we have received a maximum number of reviews overall, and the least number of reviews received are in December.



Fig:2 Graph for Number of reviews per month for the year 2018

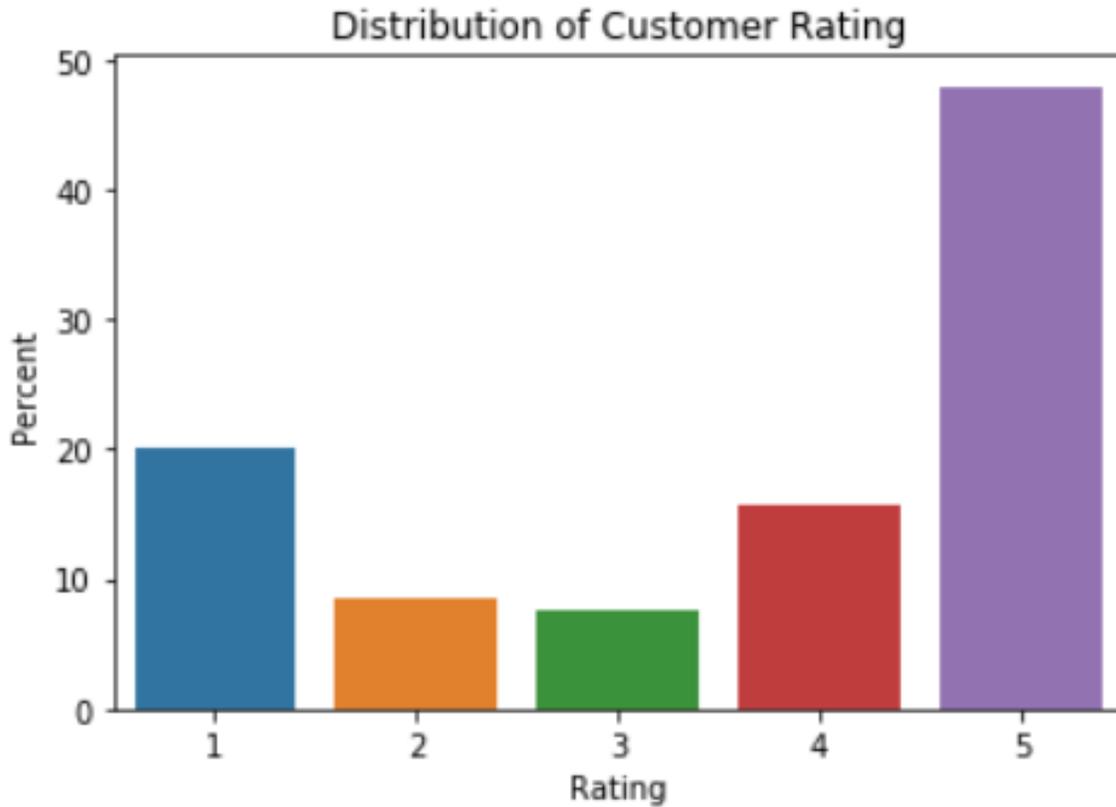


Fig:3 Graph for Distribution of Customer Rating

I analyzed the distribution of customer ratings received with a bar plot. As we can see, the maximum number of customers have given a rating of 5, and the minimum rating provided was a 3. This shows that most of the customers has a positive experience with the restaurant.

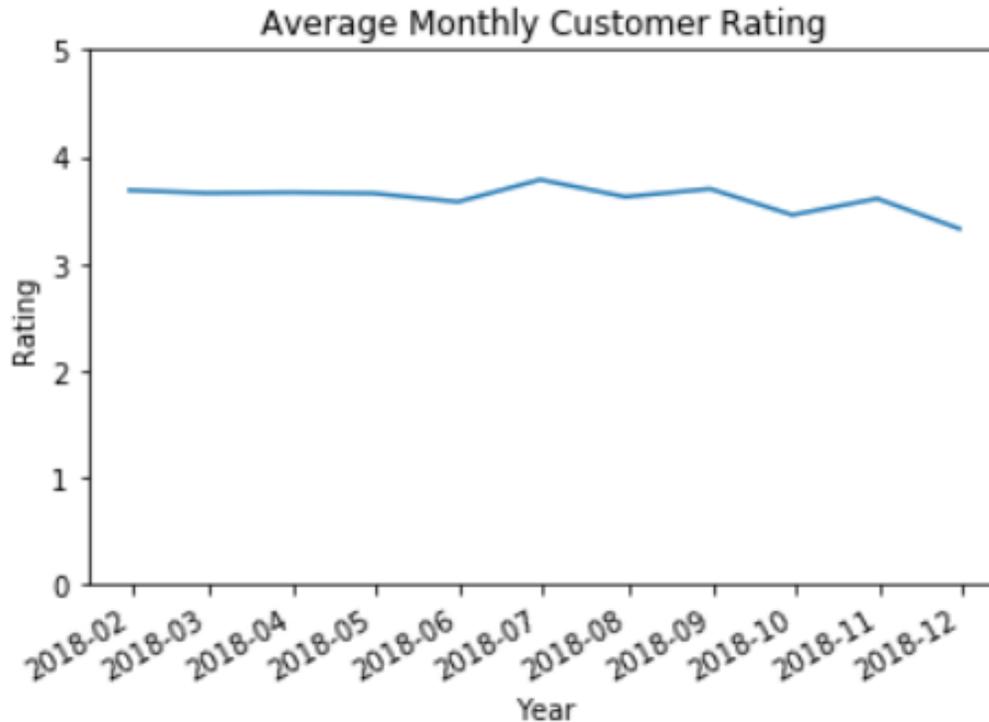


Fig:4 Graph for Average Monthly Customer Rating for the year 2018

We see that on an average most of the customers were happy with the restaurants, and above 70% of them gave ratings of 4 or 5.

3.3 Data Preprocessing

The dataset preprocessing steps include:

1. Removing reviews written language other than English.
2. Feature Selection - Selecting important columns.
3. Removing Punctuations, non-alphanumeric characters, lowercase conversion, tokenization.
4. Removing stop-words
5. Generating embeddings for each review.
6. Standardizing review-length by truncating long reviews and padding short reviews.

3.3.1 Removing reviews written in language other than English.

In this dataset, I have used only the reviews written in the English language. So, I did the language detection with a library called *langdetect* and the specific function *detect_langs*.

```
from langdetect import detect_langs

language = [detect_langs(i) for i in df.Review]
languages = [str(i[0]).split(':')[0] for i in language]
df['language'] = languages
df.language.unique()

array(['en', 'nl', 'es', 'zh-cn', 'da', 'vi'], dtype=object)
```

Date	Rating	Review	language
2018-02-15 23:15:00	1	La muchacha que nos atendió se portó muy grose...	es
2018-11-02 00:43:00	1	一次堂食，等了15分钟没人来上杯水也没人点菜。一次叫外卖，等了两小时烤鸭都做得不好。做不好菜就...	zh-cn
2018-10-01 04:41:00	1	你们猜这是什么？？这绿绿的青菜配上这些其他的不明物体和一锅汤汇成了这道著名上海菜：腌笃鲜 ...	zh-cn
2018-07-01 22:59:00	4	Solid beef brisket noodles and porknappa cabba...	da

Table: 1 List of Reviews in other languages

I have then removed these reviews that were written in different languages other than English.

3.3.2 Feature Selection - Selecting important columns

My primary focus for this project is on Reviews and Ratings, and therefore I have selected Rating, Review, and Date columns for my analysis.

```
df['Date'] = pd.to_datetime(df['Date'])
```

```
df = df[['Rating', 'Review', 'Date']]
```

```
df.head(10)
```

	Rating	Review	Date
0	1	The management of 309 is the worst i've seen. ...	2018-09-24 15:58:00
1	4	My daughter moved-in yesterday (early move-in)...	2018-08-23 13:44:00
2	1	Let me start by saying this is the single wors...	2018-11-13 17:25:00
3	2	If I didn't love the Pizza we would have stopp...	2018-05-03 12:41:00
4	1	There was terrible service and I didn't see a ...	2018-06-23 00:13:00
5	5	Pizza perfection! Although the toppings were ...	2018-05-26 14:25:00
6	5	This is definitely your casual local bar. Frie...	2018-07-08 02:49:00
7	1	AT&T Direct TV is a scam. After changing the t...	2018-09-15 17:29:00
8	1	I was treated horribly by Jason, an extremely ...	2018-03-23 21:38:00
9	5	I had walked by this vintage looking food truc...	2018-02-01 20:24:00

Table: 2 Dataset after feature selection

3.3.3 Removing Punctuations, non-alphanumeric characters, lowercase conversion, tokenization.

Table: 3 shows the reviews before removing punctuations.

2018-05-30 19:06:00	4	Awesome lunch place! The venue is not the clea...
2018-03-02 21:35:00	5	Awesome lunch. It's chipotle style where you p...
2018-04-29 23:09:00	5	Solid food at an economical price. One of the ...
2018-08-02 16:15:00	5	Best Chinese on campus. Low prices for a ton o...
2018-02-24 22:48:00	3	I frequented on a day when their chicken curry...
2018-01-10 22:39:00	1	Finally something different in Champaign! This...
2018-06-15 19:18:00	5	Stop.. Don't look for another place to eat, th...
2018-01-16 04:53:00	5	This place is outstanding and I say that as a ...
2018-10-05 15:11:00	5	Amazing food!! I highly recommend. We ordered ...
2018-01-07 16:42:00	5	Finally had a chance to visit the permanent lo...
2018-04-04 11:23:00	5	They got the best food in Champaign Illinois.....
2018-06-06 14:02:00	5	Had some food to go here and it was DEE-LISH-U...
2018-09-19 12:19:00	1	Went their once with my best friend and boyfri...
2018-06-20 18:22:00	4	Delicious! Everything I've tried here is prett...
2018-06-11 23:08:00	5	Well this Midwesterner wouldn't know "authenti...
2018-04-19 01:12:00	5	Maybe the best casual food I've ever eaten. Th...
2018-03-29 20:06:00	5	I've been here twice now, phenomenal both time...
2018-01-08 21:05:00	5	Good food at a fair price. Food has a nice spi...
2018-06-16 03:19:00	5	Surprising my EXCELLENT FOOD!!!! Wow, Champai...

Table: 3 Dataset before cleaning

Table: 4 shows reviews after removing punctuations

2018-05-30 19:06:00	Awesome lunch place The venue is not the clean...
2018-03-02 21:35:00	Awesome lunch Its chipotle style where you pic...
2018-04-29 23:09:00	Solid food at an economical price One of the b...
2018-08-02 16:15:00	Best Chinese on campus Low prices for a ton of...
2018-02-24 22:48:00	I frequented on a day when their chicken curry...
2018-01-10 22:39:00	Finally something different in Champaign This ...
2018-06-15 19:18:00	Stop Dont look for another place to eat this i...
2018-01-16 04:53:00	This place is outstanding and I say that as a ...
2018-10-05 15:11:00	Amazing food I highly recommend We ordered the...
2018-01-07 16:42:00	Finally had a chance to visit the permanent lo...
2018-04-04 11:23:00	They got the best food in Champaign Illinoisi ...
2018-06-06 14:02:00	Had some food to go here and it was DEELISHUSI...
2018-09-19 12:19:00	Went their once with my best friend and boyfri...
2018-06-20 18:22:00	Delicious Everything Ive tried here is pretty ...
2018-06-11 23:08:00	Well this Midwesterner wouldnt know authentic ...
2018-04-19 01:12:00	Maybe the best casual food Ive ever eaten The ...
2018-03-29 20:06:00	Ive been here twice now phenomenal both times ...
2018-01-08 21:05:00	Good food at a fair price Food has a nice spic...
2018-06-16 03:19:00	Surprising my EXCELLENT FOOD Wow Champaign ha...

Table: 4 Dataset after feature selection

As shown in Table: 5 we see that reviews are now converted into lowercase as “A” and “a” are not same in computer language. So, we convert all the words in reviews to lowercase.

e will awesome lunch place the venue is not the cleanest and the food can be a little too salty but cannot get a more filling delicious meal for its price anywhere on campus i come here all the time when i dont have much appetite keep in mind that its closed for a good part of the summer months awesome lunch its chipotle style where you pick rice or rice noodles with 23 toppings the food is similar to what id eat at home so i really enjoy it good amount of vegetarian options as well the tomato and egg is awesome solid food at an economical price one of the best chinese restaurants on campus also massive portions best chinese on campus low prices for a ton of food i get general taos chicken with fried rice an egg roll and a drink for under 7 i frequented on a day when their chicken curry was the lunch special naturally i followed my frugal pattern of ordering and made my request for the chicken curry it came with rice and i added mac n cheese i had very high expectations and my meal did not meet them i didnt love the flavor of the curry nor was it as spicy as i like maybe i should have asked for enhanced heat the chicken was nothing to write home about and the rice was rice the one saving component was the mac n cheese it was a delightful treat overall it was a pretty good meal but i dont think ill order the same dish again however i will keep it consistent with mac n cheese as my sidei would categorize this as a carry out establishment fo sho finally something different in champaign this was my initial thought when i found this place however this caribbean spot is just ok i was visiting family in champaign and decided to give this place a try being from chicago there are authentic jerk chicken restaurants everywhere based on the reviews on yelp i had high expectations but i was let down in my opinion it taste nothing like real jerk chicken but more of a smoked chicken that was lacking real auth

Table: 5 Reviews converted into lowercase.

3.3.4 Removing Stop-words

In this dataset, stop words are used frequently, such as “the”, “a”, “an”, “in” that add no value to the sentence. I would not want these words to take up space in our database or taking up valuable processing time. So, it is necessary to remove the irrelevant data and keep the most relevant strings and words from the reviews. Hence, I have filtered stop words from the text with a corpus library and the specific function *stopwords*. NLTK(Natural Language Toolkit) in python has a list of stopwords stored in 16 different languages.

```
from nltk.corpus import stopwords
from sklearn.feature_extraction.stop_words import ENGLISH_STOP_WORDS

my_stop_words = set(stopwords.words('english') + list(ENGLISH_STOP_WORDS))
```

I then wanted to investigate the words with the highest frequency that customers have used in reviews. Word cloud is a smart way to highlight words with high-frequency in textual data. The more a specific word appears in the text, the bigger and bolder it will be in the word cloud.



Fig: 5 Word Cloud for highest frequency words

3.3.5 Generating embeddings for each review.

The model cannot accept the text data directly as it first needs to be converted from an array of words to vector values. Keras provides a Tokenizer module that encodes the texts. First, it tokenizes sentences into words. After converting the texts to sequences, we obtain the word_index to all tokenized words. I have then used texts_to_sequences to convert the tokenized words into a sequence that encodes the words into numbers as shown in Fig: 6.

```
{'the': 1, 'and': 2, 'to': 3, 'a': 4, 'i': 5, 'was': 6, 'of': 7, 'is': 8,
'for': 9, 'in': 10, 'it': 11, 'my': 12, 'that': 13, 'with': 14, 't': 15, '
this': 16, 'on': 17, 'they': 18, 'have': 19, 'you': 20, 'had': 21, 'not':
22, 's': 23, 'are': 24, 'at': 25, 'were': 26, 'we': 27, 'be': 28, 'food':
29, 'me': 30, 'as': 31, 'very': 32, 'so': 33, 'place': 34, 'out': 35, 'but
': 36, 'good': 37, 'their': 38, 'great': 39, 'get': 40, 'time': 41, 'would
': 42, 'if': 43, 'just': 44, 'here': 45, 'our': 46, 'all': 47, 'an': 48, '
from': 49, 'there': 50, 'like': 51, 'service': 52, 'one': 53, 'up': 54, 'c
an': 55,
```

Fig: 6 Tokenized words with word index

After this step, I have generated encoded reviews by iterating through the texts and mapping each word to its embedding value by an Embedding layer. Each word is then mapped to a real-valued vector. To eliminate the variance is to replace the embeddings. In standard embeddings, words that are semantic synonyms have unrelated vector representations. So, the words' such as "Kitchen" and "Dinner" could be very far away from each other, while they should be close since their meaning shares a theme. I have used the GloVe encoding method that encodes values that give words in the same space closer to vectors compared to distinct words. Various things are occurring here, the first of which is reading the glove embeddings. I have read them from the text file, split them into the word and the vector. With the help of mapping words and vectors, I have created an Embeddings Matrix that maps all the words in *our* vocab to the vector for it. Iterating over it, I have built the embeddings matrix by indexing the glove vectors.

```
[[ 0.          0.          0.          ... 0.          0.
  0.          ]
 [-0.038194 -0.24487001  0.72812003 ... -0.1459      0.82779998
  0.27061999]
 [-0.071953  0.23127      0.023731   ... -0.71894997  0.86894
  0.19539    ]
 ...
 [ 0.          0.          0.          ... 0.          0.
  0.          ]
 [ 0.1569      0.55396998 -0.20557     ... 0.32764      0.36052001
 -0.087964   ]
 [ 0.          0.          0.          ... 0.          0.
  0.          ]]
```

Fig: 7 GloVe embedding matrix

3.3.6 Standardizing review-length by truncating long reviews and padding short reviews.

All of the reviews have different lengths, some have 50 words, some have 100, and some are even larger. We can take a look at a histogram of the review lengths to make the decision on where to cap the reviews. In this step, I have standardized the reviews; it was observed that 20–40 words length is the most common length for the reviews. Even a few words can convey if a review is positive or negative. So, to save processing power and reduce bias between review lengths, standardization of a fixed length for all reviews is achieved using `pad_sequences` from Keras.

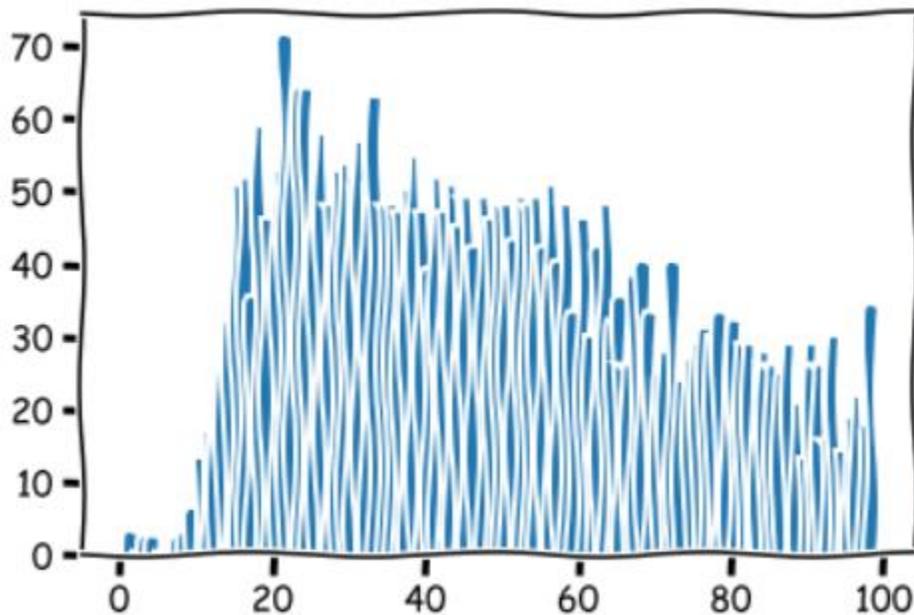


Fig: 8 Histogram of Review Lengths

3.4 Selection of Machine Learning Models

The Yelp dataset that I have chosen is an excellent source for the quality of reviews from consumers about a restaurant. I have built an LSTM model and trained the data on this model to classify the reviews as “positive” and “negative”. Further, I have checked the performance of the model on test data. I have also used libraries for different classifier models such as Naïve Bayes, Bernoulli, Logistic Regression, Support Vector Machine, Stochastic Gradient Descent, and Long Short Term and trained the data on all these models and tested the performances of these models on test data. I have taken the Naïve Bayes model as my baseline model and then compared all the results to identify the model that best predicts and classifies the reviews into “positive” and “negative”.

3.4.1 Naïve Bayes classifier

Based on Bayes’ Theorem, Naive Bayes classifiers are a collection of classification algorithms. It is not a unique algorithm but a group of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other. It is the simplest form of Bayesian Network, in which all attributes are independent given the value of the class variable. Hence, it is called conditional independence. It assumes each feature is conditional independent to other features given the class. A Naive Bayes classifier is a technique that applies to a specific class of problems, namely those that are phrased as associating an object with a discrete category. Naive Bayes has several advantages from the numerical-based approach group, such as simple, fast, and high accuracy.

3.4.2 Bernoulli Classifier

Bernoulli Classifier is an extension of the Naive Bayes Classifier, and it is equivalent to the binary independence model, which generates an indicator for each term of vocabulary. This model is similar to the multinomial Naive Bayes, but the predictors are boolean variables. We use the parameters to predict the class variable that considers only values yes or no, for example, if a word is present in the text or not.

3.4.3 Logistic Regression

Although many complex extensions exist, Logistic regression, a statistic model in its primary form, utilizes a logistic function to model a binary dependent variable. In **regression** analysis, **logistic regression** (or **logit regression**) estimates a logistic model's parameters (a form of binary **regression**). When the dependent variable(target) is categorical, we use Logistic regression.

For example,

- To predict whether the transaction was fraudulent (1) or (0)
- Whether the person is happy (1) or not (0)

It is used for the classification of the user review in positive and negative classes.

3.4.4 Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent (SGD) Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to discriminative learning of linear classifiers under convex loss functions such as (linear) Support vector Machines and Logistic Regression. Even though SGD has been around in the machine learning community for a long time, it has received a considerable amount of attention just recently in the context of large-scale learning.

3.4.5 Support Vector Machine (SVM)

The support vector machine algorithm's objective is to find a hyperplane in N-dimensional space where N is the number of features that distinctly classify the data points. Support vectors are data points that are closer to the hyperplane and influence the hyperplane's position and orientation. The main objective is to find a plane with the maximum margin, i.e., the maximum distance between both classes' data points. Using these support vectors, we maximize the classifier's margin that provides some reinforcement to classify data points with more confidence. Support vector Machines use a Support vector classifier for the classification of the user review in the classes. Using this method is efficient for high dimensional spaces and effective when the number of dimensions is greater than the number of samples.

3.4.6 Long Short Term Memory (LSTM- a deep learning approach)

LSTM is a recurrent neural network (RNN) that is used in the field of deep learning. Unlike feed-forward neural networks, it has feedback connections, i.e., it can process single data points and a sequence of data. It is required in complex problem domains like machine translation, speech recognition, and more. Therefore, LSTM can be used for sentiment analysis.

3.4.7 Designing LSTM Model

Post data preparation, I have designed a model and then trained it to predict accurate results. After comparing LSTM and RNN, I found that RNNs have feedback loops in the recurrent layer. Still, it is challenging to train standard RNNs to solve problems requiring long-term temporal dependencies as it suffers from vanishing gradient problems. LSTM networks are a variety of RNN that utilizes special units in addition to standard units that include a 'memory cell' that can maintain information in memory for a more extended period. It deals with these problems by introducing new gates, such as input and forget gates, which allow for better control over the gradient flow and enable better preservation of "long-range dependencies.

3.4.7.1 Models and Hyperparameters Tuning

3.4.7.1.1 LSTM model:

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 100, 100)	3852000
lstm_2 (LSTM)	(None, 100, 70)	47880
dropout_1 (Dropout)	(None, 100, 70)	0
lstm_3 (LSTM)	(None, 70)	39480
dense_2 (Dense)	(None, 100)	7100
dense_3 (Dense)	(None, 2)	202
Total params: 3,946,662		
Trainable params: 3,946,662		
Non-trainable params: 0		

Fig: 9 LSTM Model

As a part of my project, I have designed an LSTM based neural network with

- an embedding layer that helps convert each word into a vector.
- 2 LSTM layers,
- a dropout layer is used for regularization,
- two dense layers; as a dense layer is a fully connected **layer**, all the neurons in a **layer** are connected to those in the next **layer**.

Trained the model and performed predictions on test datasets. Post the training testing phase; the results will be compared to other classification model's performance to see which model gives more accurate results.

3.4.7.1.2 Hyperparameters

Relu – The **rectified linear activation function** or **ReLU** is a piecewise linear function that will output the input directly if it is positive. Otherwise, it will output zero.

Softmax – It is an activation function that turns logits into probabilities that sum to one. Softmax function outputs a vector that represents the probability distributions of a list of potential outcomes.

Adam – It computes individual learning rates for different parameters.

Binary Cross-Entropy – It is also called Sigmoid **Cross-Entropy** loss. It is a Sigmoid activation plus a **Cross-Entropy** loss. It is independent for each vector component, which means the loss computed for every CNN output vector component is not affected by other component values.

Number of Epochs	6
Activations Functions	Relu and softmax
Optimizer	Adam
Loss Function	Binary crossentropy
Learning Rate	Learning Rate = 0.001, Decay = 0.0001
Batch Size	64

Table: 6 Summarized view of hyperparameters

4. Evaluation of Models

4.1 Evaluation Metrics

Precision, Recall and F1-Score are used to measure the accuracy of the classification model.

1. Precision – Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It tells us how precise the model is out of those predicted positive, how many of them are actual positive values. If the model avoids lot of mistakes in predicting correct values then the model has high precision.
2. Recall – Recall is the ratio of correctly predicted positive observations to all observations in the actual class.
3. F1 Score – F1 Score is the weighted average of Precision and Recall.
4. Accuracy – Accuracy is the measure of correctness of the value in correlation with the information.

4.2 Results of different classification model

While working with different models, its easy to see the contrast when training models. Naïve Bayes is considered as a baseline model. From the below-tabulated data, the results i.e., Evaluation Matrics of Naïve Bayes model, are shown. These results are then compared to other classification models result and identify which classifier provides more accurate results.

4.2.1 Naïve Bayes Classifier

Evaluation Matrics	Naïve Bayes
Precision	0.8492
Recall	0.9147
Accuracy	0.8824
F_1 score	0.8285

Table: 7 Table for sentiment analysis using Naïve Bayes Classifier

As shown in Fig: 10, we see that the Recall is high, and precision is the least for this model.

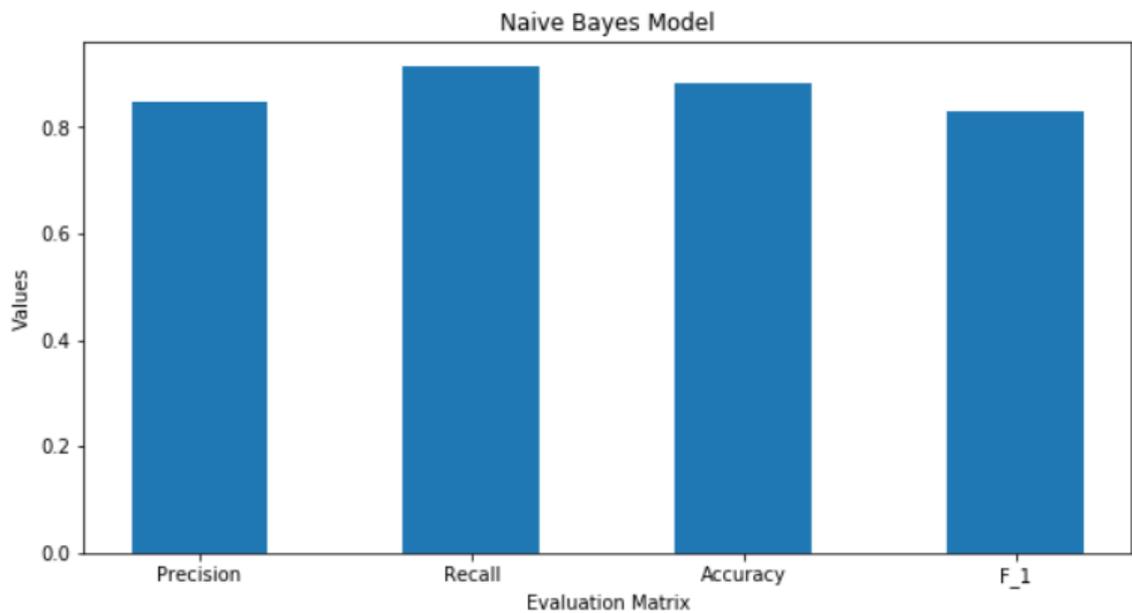


Fig: 10 Graph for sentiment analysis using Naïve Bayes Classifier

4.2.2 Bernoulli Classifier

Evaluation Metrics	Bernoulli's Classifier
Precision	0.8297
Recall	0.9324
Accuracy	0.8781
F_1 score	0.8207

Table:8 Table for sentiment analysis using Bernoulli Classifier

As shown in Fig: 11, we see that the Recall is high, and F_1 is the least for this model.

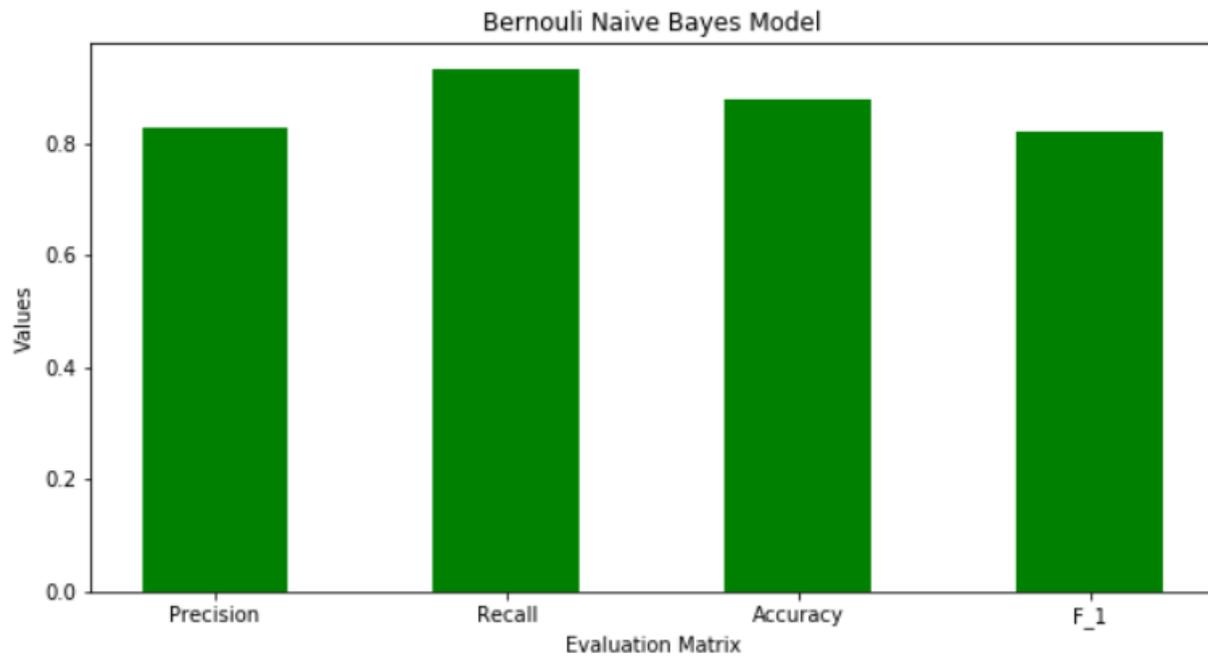


Fig: 11 Graph for sentiment analysis using Bernoulli Classifier

4.2.3 Logistic Regression

Evaluation Metrics	Logistic Regression
Precision	0.8951
Recall	0.9469
Accuracy	0.9221
F_1 score	0.8864

Table:9 Table for sentiment analysis using Logistic Regression

As shown in Fig: 12, we see that the Recall is high, and F_1 is the least for this model.

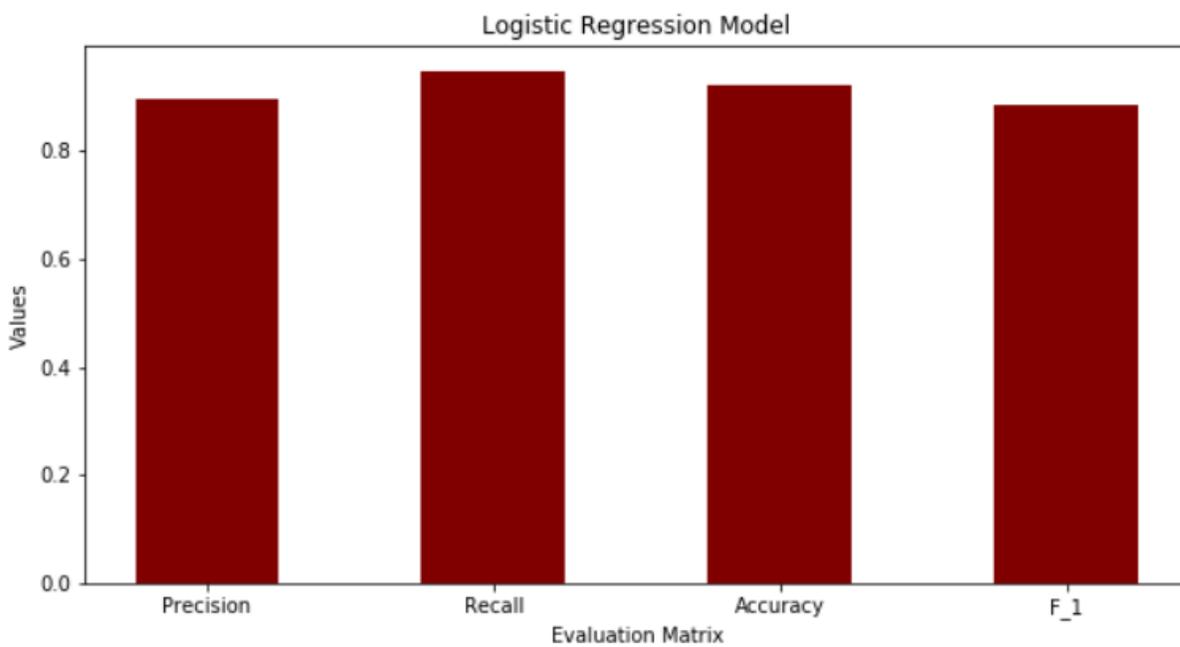


Fig: 12 Graph for sentiment analysis using Logistic Regression

4.2.4 Stochastic Gradient Descent

Evaluation Metrics	SGD
Precision	0.8561
Recall	0.8796
Accuracy	0.8304
F_1 score	0.8676

Table:10 Table for sentiment analysis using Stochastic Gradient Descent

As shown in Fig: 13, we see that the Recall is high, and accuracy is the least for this model.

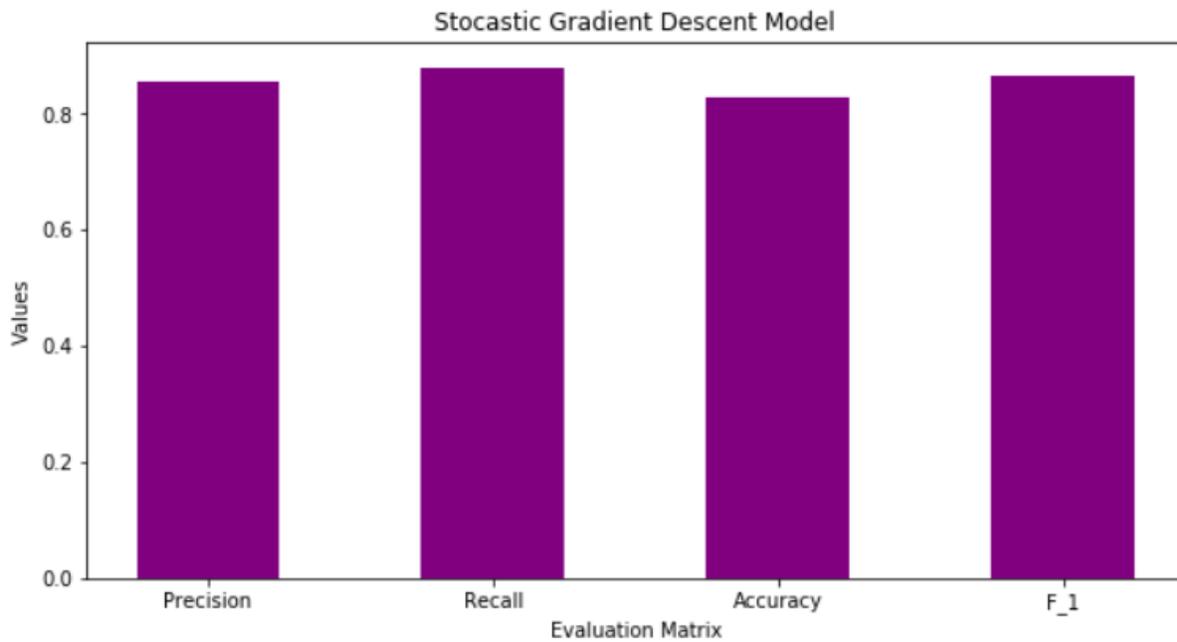


Fig: 13 Graph for sentiment analysis using Stochastic Gradient Descent

4.2.5 Support Vector Machine

Evaluation Metrics	SVM
Precision	0.9212
Recall	0.9228
Accuracy	0.8471
F_1 score	0.9022

Table: 11 Table for sentiment analysis using Support Vector Machine

As shown in Fig: 14, we see that the precision is high, and accuracy is the least for this model.

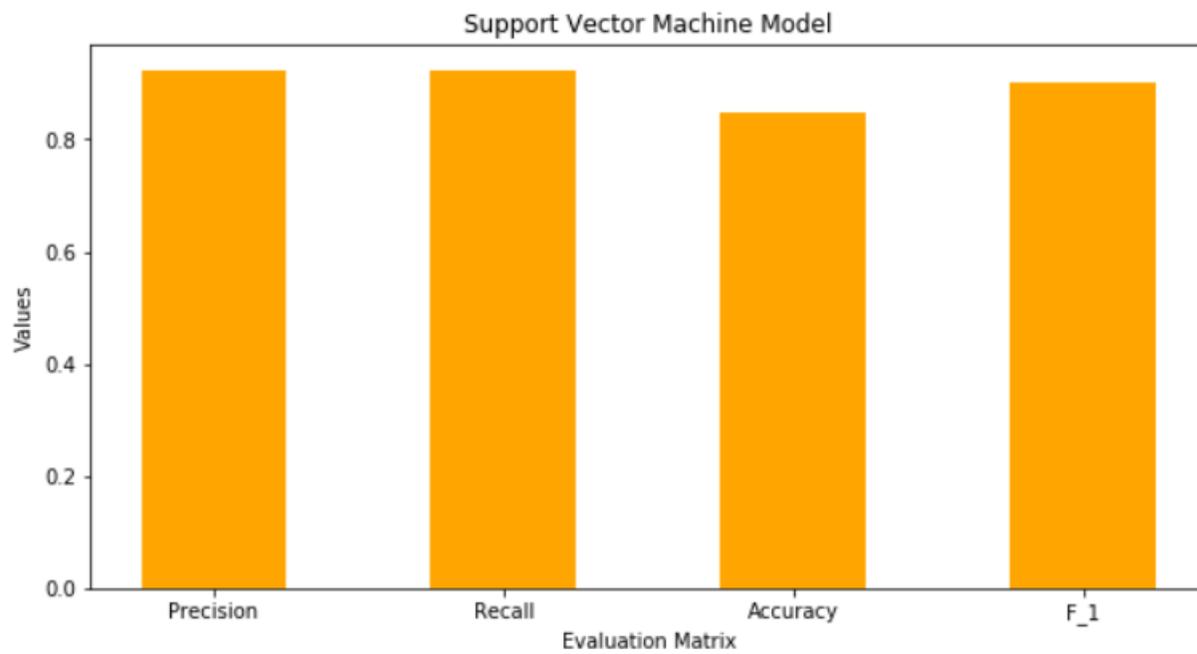


Fig: 14 Graph for sentiment analysis using Support Vector Machine

4.2.6 LSTM

Evaluation Metrics	LSTM
Precision	0.807
Recall	0.8905
Accuracy	0.847
F_1 score	0.9021

Table: 12 Table for sentiment analysis using LSTM

As shown in Fig: 15, we see that the F_1 is high, and precision is the least for this model.

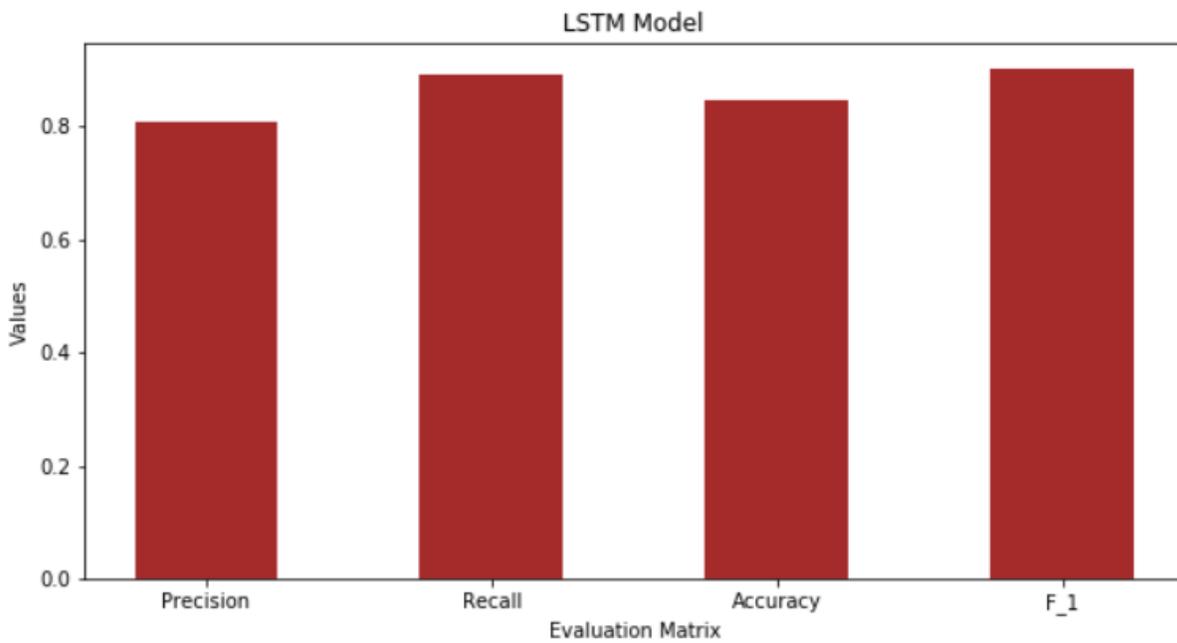


Fig: 15 Graph for sentiment analysis using LSTM

5. Comparing Evaluation Scores

5.1 Summarized Results for classification models

The table summarizing the precision, recall, accuracy and F1 Score, of all the models trained on dataset is tabulated below.

	Naïve Bayes	Bernoullis	Logistic Regression	SGD	SVM	LSTM
Precision	0.8492	0.8297	0.8951	0.8561	0.9212	0.8070
Recall	0.9147	0.9324	0.9469	0.8796	0.9228	0.8905
Accuracy	0.8824	0.8781	0.9221	0.8304	0.8471	0.8470
F_1 score	0.8285	0.8207	0.8864	0.8676	0.9022	0.9021

Table: 13 Table for analysis of evaluation values of all models

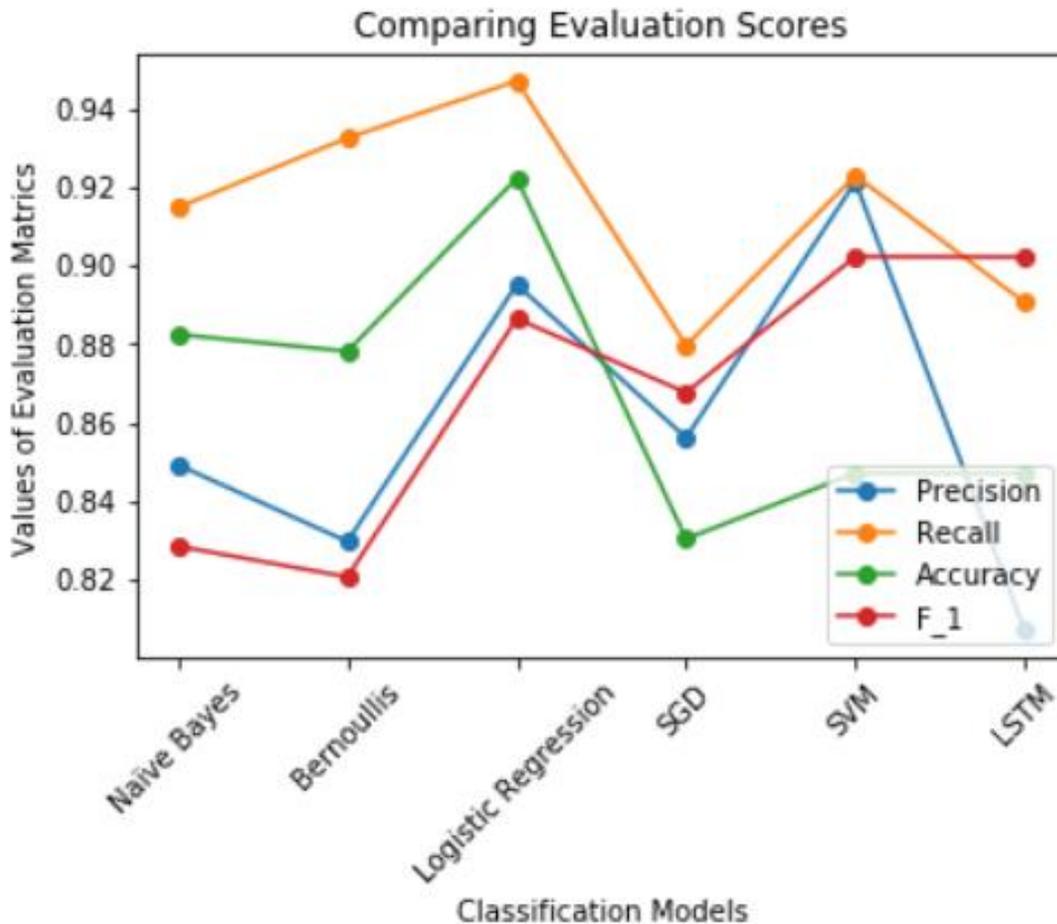


Fig: 16 Comparison of Evaluation Scores

Precision is the highest in SVM compared to all other classifier models whereas lowest in LSTM. Recall is high in Bernoulli's and low in LSTM. F1 score is the highest in LSTM and lowest in Bernoulli's model. Accuracy is high in SVM and low in LSTM.

6. Conclusions

I have worked on classifying reviews into positive, negative, and neutral. Further, I have used different machine learning models and compared their performance in classifying positive and negative reviews. I have considered high precision as it contains fewer false positives and gives a more accurate prediction. SVM model provides the highest precision with fewer false positives compared to the performance of other classification models. Hence, I have concluded SVM to be the best model for classifying the reviews to either positive or negative.

References

- [1] Lina L. Dhande and Dr. Prof. Girish K. Patnaik, "Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier", IJETTCS, Volume 3, Issue 4 July-August 2014, ISSN 2278-6856.
- [2] P.Kalaivani, "Sentiment Classification of Movie Reviews by supervised machine learning approaches" et.al, Indian Journal of Computer Science and Engineering (IJCSE) ISSN : 0976-5166 Vol. 4 No.4 Aug-Sep 2013.
- [3] Meena Rambocas, João Gama, "Marketing Research: The Role of Sentiment Analysis", April 2013, ISSN: 0870-8541.
- [4] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg, 2005.
- [5] "Movie review dataset," [Online]. Available <http://www.cs.cornell.edu/people/pabo/movie-reviewdata/>, [Accessed: October 2013].
- [6] K. M. Leung, "Naive Bayesian classifier," [Online] Available: <http://www.sharepdf.com/81fb247fa7c54680a94dc0f3a253fd85/naiveBayesianClassifier.pdf>, [Accessed: September 2013].
- [7] Zhou Yong , Li Youwen and Xia Shixiong "An Improved KNN Text Classification Algorithm Based on Clustering", journal of computers, vol. 4, no. 3, march 2009.
- [8] B. Pang, L. Lee, S. Vaithyanathan, "sentiment classification using machine learning techniques," In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 79-86. July 2002.

- [9] S. Deng, T. Mitsubuchi, K. Shioda, T. Shimada, A. Sakurai, “Combining technical analysis with sentiment analysis for stock price prediction,” in Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2011, pp. 800- 807.
- [10] K. M. Leung, “Naive Bayesian classifier,” [Online]. Available: <http://www.sharepdf.com/81fb247fa7c54680a94dc0f3a253fd85/naiveBayesianClassifier.pdf>, [Accessed: September 2013].
- [11] Y. Qiu, G. Yang, and Z. Tan, “Chinese text classification based on extended nave bayes model with weighed positive features,” in First International Conference on Pervasive Computing, Signal.
- [12] https://en.wikipedia.org/wiki/Logistic_regression
- [13] Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, Sweta Tiwar , “Sentiment Analysis of Review Datasets using Naïve Bayes’ and K-NN Classifier” <https://arxiv.org/ftp/arxiv/papers/1610/1610.09982.pdf>
- [14] Boya Yu, Jiaxu Zhou, Yi Zhang, Yunong Cao , “Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews” <https://arxiv.org/ftp/arxiv/papers/1709/1709.08698.pdf>
- [15] Bing Liu, “Sentiment Analysis and Opinion Mining” <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.244.9480&rep=rep1&type=pdf>