International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IJIEMR Transactions, online available on 26th Dec 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=ISSUE-12

DOI: 10.48047/IJIEMR/V11/ISSUE 12/49

Title **EXTENSIVE STUDY OF EXPERT SYSTEM FOR RESTAURANT SELECTION BASED ON SENTIMENTAL ANALYSIS**

Volume 11, Issue 12, Pages: 369-374

Paper Authors
Dr.G.Padmaja, Dr.K.S.R.Radhika

EXTENSIVE STUDY OF EXPERT SYSTEM FOR RESTAURANT SELECTION BASED ON SENTIMENTAL ANALYSIS

[1] Dr.G.Padmaja, [2] Dr.K.S.R.Radhika

[1] Professor, Potti Sriramulu Chalavadhi Mallikarjunarao College of Engineering & Technology, Vijayawada, Andhra Pradesh. E-Mail: padmajagrandhe@gmail.com

[2] Professor, TKR College of Engineering & Technology, Hyderabad, Telangana. E-Mail: kammilisrr@gmail.com

Abstract: In this digital world, every interactions of the user with internet are being recorded. The systems which provide recommendations based on the user past events are getting popularity to grab the attention of the user and to make the busy life of the user easier. An adaptive based recommendation system is proposed and is compared with standard recommendation algorithms. The proposed system considers the reviews as base factor and helps the users to find an appropriate restaurant among the various multi cuisines with in and around locations. The proposed system considers the context of the reviews for designing an efficient model.

Keywords: Pos Tagging, Sentiment Analysis, Machine Learning, NLP

1. Introduction

Recommendation systems make the life of users easier by analyzing their behavioral patterns and suggesting the things based on their interests. Recommendation system plays an important role in the process of customizing the users preferences in the digital world. It creates an platform for the E-commerce sites to grow their business by promoting ads or mails or messages by filtering the data about the users which are collected from data sources. The popular mechanisms in the recommendation systems are: Collaborative filtering, it collects reviews about a product from different people and performs opinion mining to decide whether to recommend a product or not. The main advantage of the collaborative filtering is it generates efficient system than the other type of filtering in terms of accuracy. The drawback is, it is inefficient if the system lacks the previous data. Content filtering, is an another popular mechanism, which concentrates on a single user rather than multiple users. These mechanisms are good at case based reasoning process, the more data it collects from the user, the more accuracy it can generate for the system. Simultaneously, it may suffer with cold start problem, which arises because of insufficient data from the user. It generally occurs with the users who are new to the environment.

II. Literature Review:

Geetika Gautam [2] designed a system to analyze the twitter dataset. The dataset contains data from different sources, so to reduce the inconsistencies among the data, the system polarizes the data and unigram acts as feature extraction technique for identifying the adjectives in the sentences. A modified Naïve Bayesian algorithm is proposed, which learns from the classified data and assigns the document to the correct tweet classifier. All the features extracted are added back to the preprocessed dataset and synonyms are identified to polarize the data. The model calculates the similarity for synonyms and generate the reviews.

Milan Gaonkar [6] proposed a novel algorithm "Sentidetect", using the dataset from amazon product review system. In this model, all the opinion words are gathered based on the adjectives then polarization is performed using AFINN dictionary, which assigns the word score in between -5 to +5. The opinion bearing technique divides the words into True-Positive, True-Negative, False-Positive and False-Negative categories. The goal of the system is to make a user to decide whether to buy a product or not in a informed approach. The advantage of this model is it uses the supervised machine learning algorithms for partitioning the dataset into training and testing.

Rasika Wagh [9] developed 3 supervised classification algorithms on twitter dataset. The developed model performs preprocessing by removing the retweets, special and punctuation marks, tokenization and stemming. In the next step, polarity of the words is done with the help of predefined lexicons. The semantic detection finds the category of the word as either positive , negative or neutral. A hybrid approach is applied , which improved the accuracy by 4-5% than the traditional algorithms.

Aliza Sarlan [1] proposed system using twitter dataset on customer review using machine learning approach. The system first computes the POS tagging for each twitter post and a document-word vector is constructed based on the Top-N adjective words by navigating through out the dataset. The TF-IDF matrix is calculated by finding occurrence of each word in all the documents. Lexicon approach calculates the polarity of the sentence and then the accuracy of the model is computed.

Lopamudra Dey [4] designed Naïve Bayesian and KNN algorithms on the movie reviews and hotel reviews datasets. Naïve Bayesian need works efficiently even with small amount of data. The Naïve Bayesian algorithm divides the data into positive reviews and negative reviews. This model does tag description to each sentence and it calculates the conditional probability score for each word and finds the best words among them. The KNN algorithm calculates the scores of the word by applying chi square test on the training dataset and algorithm adds nearest neighbor score to the calculated score. Finally the model generates the review score and performs the evaluation.

III Proposed System:

Data collection: The data contains customer satisfaction reviews with two columns. The second column act as class label, to represent whether the review is positive or negative. The dataset uses 1 to represent positive review and 0 to represent negative. The dataset contains 1000 customers' satisfaction reviews, to improve the quality of the restaurant. The data is collected from the Kaggle, open source repository of datasets [11]. The workflow of the proposed system is shown in the figure 1.

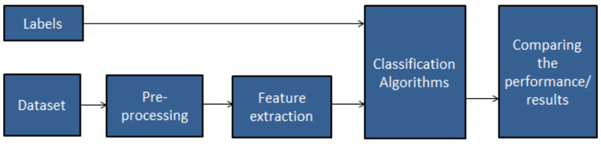


Figure 1: Proposed System Architecture

- **Pre-processing:** The customer reviews are categorical data, to process these types of data; the proposed system uses

Natural Language Processing. The Natural Language Understanding (NLU), extracts the information from the different sources and performs the sentiment analysis, which helps to separate the positive part and negative part of the sentence and it calculates the polarity of the sentence, to assign the sentiment score. The sentiment score determines whether it is a positive review or negative review. To pre-process this categorical data the proposed system performs the following subsection.

Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing .These stream of chunks helps in understanding the context of the sentence. During this process, the proposed system also takes care about stop word removal process, because these act as sort of noisy data, reduction of these words can fasten the process as well as the model can extract more accurate information. Stop words are most frequent words that occurs in English language like a, an, the, is, for and etc.,

- **Feature extraction:** To convert the raw data into matrix format, the proposed system discusses about some feature extraction techniques in the below sub sections:

Bag-of-Words: The generated matrix consists of vector representations, these vectors are partitioned into training and test datasets. This mechanism searches for the occurrence of the known words in a document, it just concerned about the presence and absence of the known words. The mechanism identifies all unique words in the given sentence by ignoring the special characters and punctuation symbols, and then it creates a document vector containing the number of occurrences of each word in the sentence.

Construction of N-grams: The N-gram technique calculates the probability with different sequences of sentences i.e., the occurrence of the word among the previously stored words is calculated. To reduce the time complexity of the model, the proposed system implements **Markov assumption, which assumes the occurrences by considering the last word of the sentence.**

- **Splitting into Training and Testing data:** Machine Learning algorithms calculates the accuracy of the system by dividing the data into two sets. The thumb rule of the splitting the data is no training values should be included into test data, because inclusion of which may lead to overfitting problem. So, the proposed system splits the data as follows:

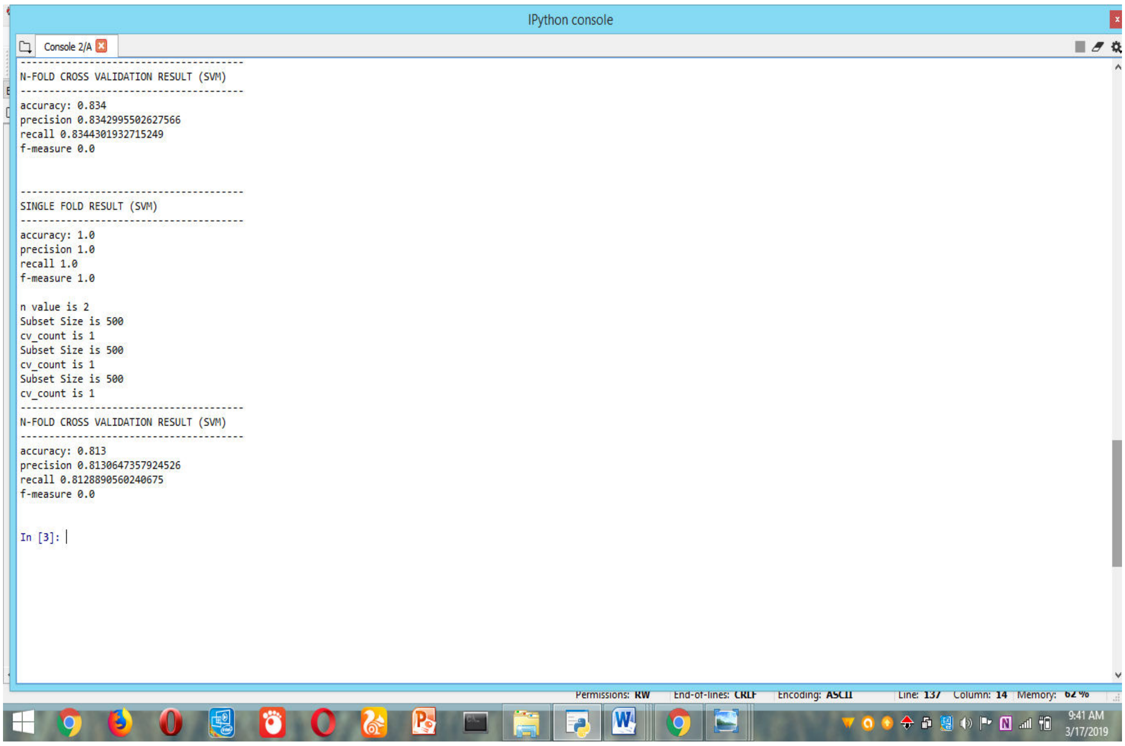
Train set = 75% of positive_data + 75% of negative_data

Test set = 25% of positive_data + 25% of negative data

- **Classification Algorithms:**

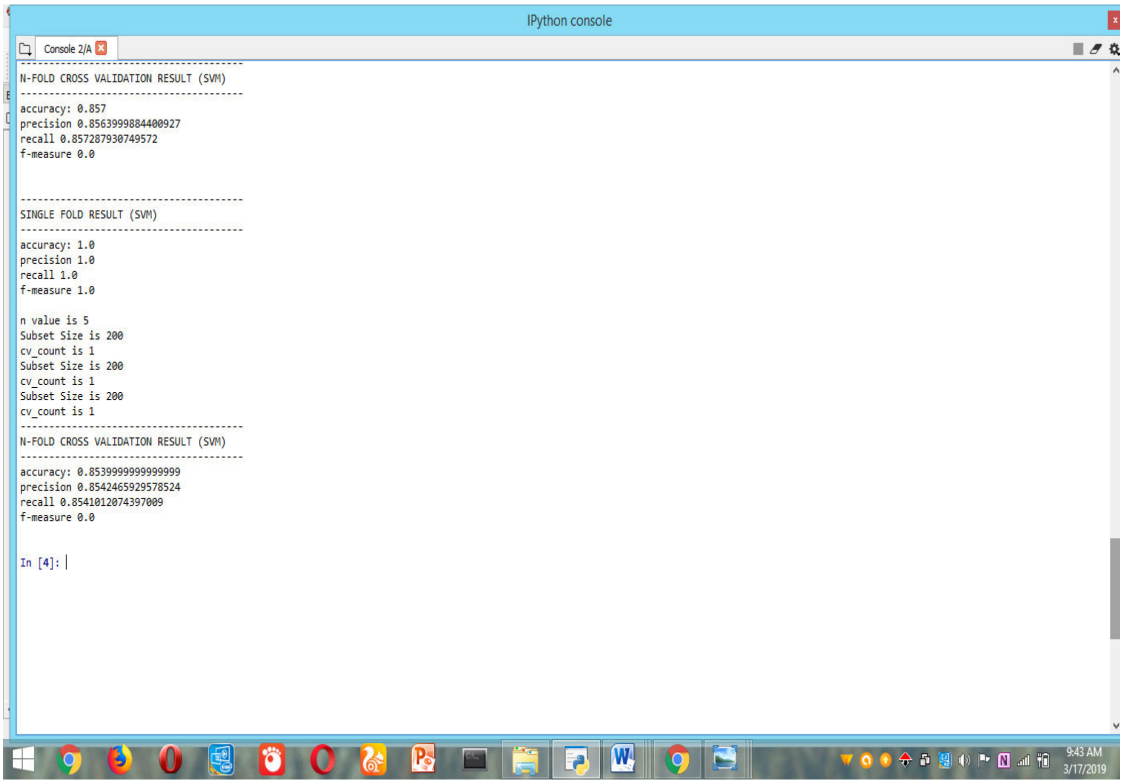
The proposed system, in order to classify the reviews based on sentiment we have used three Supervised Machine learning algorithms such as Naïve Bayes Algorithms, Support Vector Machine, Maximum Entropy algorithms. These Supervised Machine algorithms have been used for classification problems using the labels which we will provide. By using the feats and the predefined labels for data, these supervised machine learning algorithms will be trained. Later these algorithms will be tested on testing data using which these algorithms will predict the labels for the respective feats present in the test data.

IV. Results:



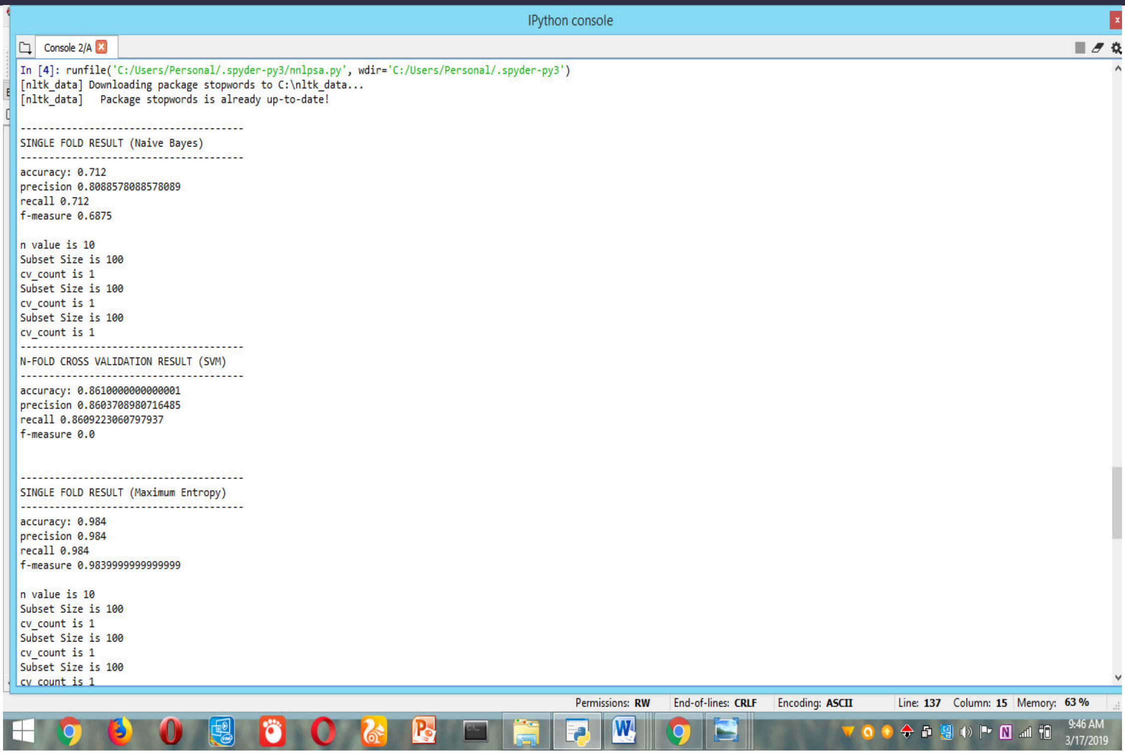
```
----------------------------------------
N-FOLD CROSS VALIDATION RESULT (SVM)
----------------------------------------
accuracy: 0.834
precision 0.8342995502627566
recall 0.8344301932715249
f-measure 0.0


----------------------------------------
SINGLE FOLD RESULT (SVM)
----------------------------------------
accuracy: 1.0
precision 1.0
recall 1.0
f-measure 1.0

n value is 2
Subset Size is 500
cv_count is 1
Subset Size is 500
cv_count is 1
Subset Size is 500
cv_count is 1
----------------------------------------
N-FOLD CROSS VALIDATION RESULT (SVM)
----------------------------------------
accuracy: 0.813
precision 0.8130647357924526
recall 0.8128890560240675
f-measure 0.0


In [3]: 
```

Fig 1: Results of Classification algorithms when n=2



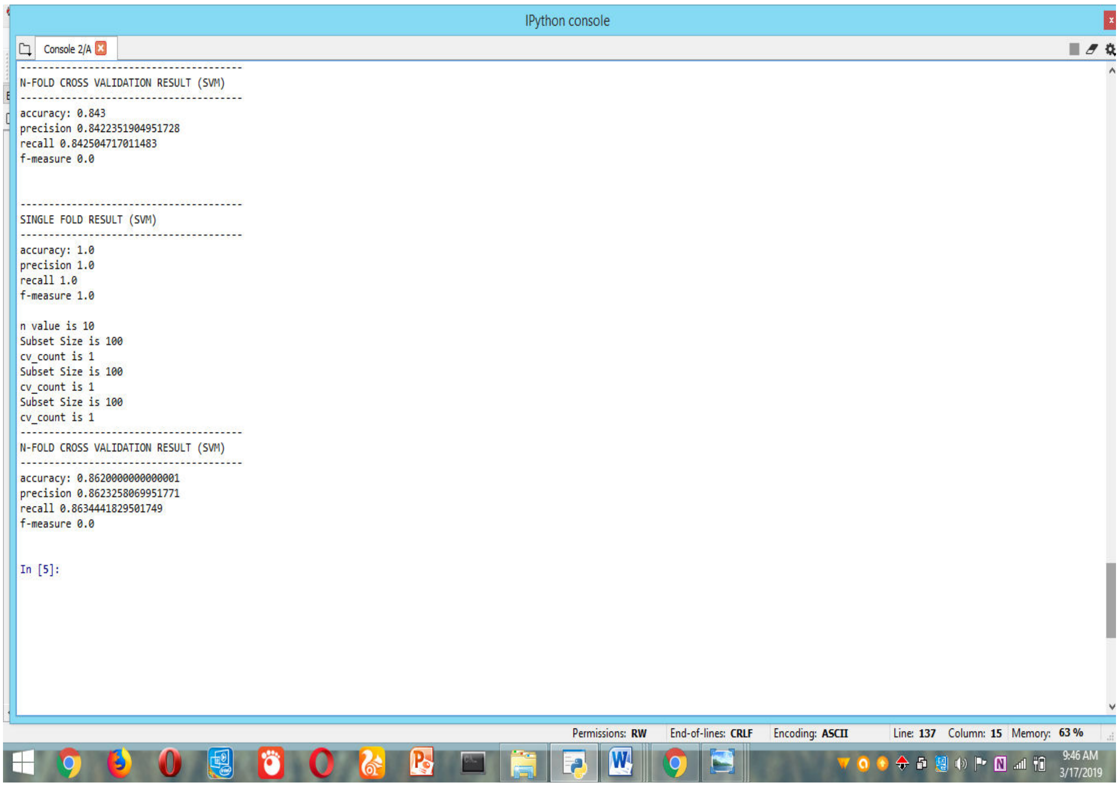
```
----------------------------------------
N-FOLD CROSS VALIDATION RESULT (SVM)
----------------------------------------
accuracy: 0.857
precision 0.8563999884400927
recall 0.857287930749572
f-measure 0.0


----------------------------------------
SINGLE FOLD RESULT (SVM)
----------------------------------------
accuracy: 1.0
precision 1.0
recall 1.0
f-measure 1.0

n value is 5
Subset Size is 200
cv_count is 1
Subset Size is 200
cv_count is 1
Subset Size is 200
cv_count is 1
----------------------------------------
N-FOLD CROSS VALIDATION RESULT (SVM)
----------------------------------------
accuracy: 0.8539999999999999
precision 0.8542465929578524
recall 0.8541012074397009
f-measure 0.0


In [4]: 
```

Fig 2: Results of Classification algorithms when n=5



```
In [4]: runfile('C:/Users/Personal/.spyder-py3/nnlpsa.py', wdir='C:/Users/Personal/.spyder-py3')
[nltk_data] Downloading package stopwords to C:\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!

----------------------------------------
SINGLE FOLD RESULT (Naive Bayes)
----------------------------------------
accuracy: 0.712
precision 0.8088578088578089
recall 0.712
f-measure 0.6875

n value is 10
Subset Size is 100
cv_count is 1
Subset Size is 100
cv_count is 1
Subset Size is 100
cv_count is 1
----------------------------------------
N-FOLD CROSS VALIDATION RESULT (SVM)
----------------------------------------
accuracy: 0.8610000000000001
precision 0.8603708980716485
recall 0.8609223060797937
f-measure 0.0


----------------------------------------
SINGLE FOLD RESULT (Maximum Entropy)
----------------------------------------
accuracy: 0.984
precision 0.984
recall 0.984
f-measure 0.9839999999999999

n value is 10
Subset Size is 100
cv_count is 1
Subset Size is 100
cv_count is 1
Subset Size is 100
cv_count is 1
```

Fig 3: Results of Classification algorithms when n=10



```
----------------------------------------
N-FOLD CROSS VALIDATION RESULT (SVM)
----------------------------------------
accuracy: 0.843
precision 0.8422351904951728
recall 0.842504717011483
f-measure 0.0


----------------------------------------
SINGLE FOLD RESULT (SVM)
----------------------------------------
accuracy: 1.0
precision 1.0
recall 1.0
f-measure 1.0

n value is 10
Subset Size is 100
cv_count is 1
Subset Size is 100
cv_count is 1
Subset Size is 100
cv_count is 1
----------------------------------------
N-FOLD CROSS VALIDATION RESULT (SVM)
----------------------------------------
accuracy: 0.8620000000000001
precision 0.8623258069951771
recall 0.8634441829501749
f-measure 0.0


In [5]:
```

Fig 4: Results of Classification algorithms when n=10

Conclusion:

The proposed system aims to classify tweets, it can help us to know reviews of people about a restaurant. Also, the main concern in Machine Learning Algorithm is the accuracy of the result. So our proposed system has compared the accuracies of various machine learning algorithms using n-fold cross validation results. As a result, the accuracy of our model also increases which is our main concern.

As we have seen that this model can be used to analyze sentiment, if we feed model with the tweets regarding Restaurant and we find out that most of the tweets are showing "Positive" that means the Restaurant can be recommended.

References:

[1] Aliza Sarlan,Shuib Basri,Chayanit Nadam."Twitter sentiment analysis." IEEE 2014

[2] Geetika Gautam, Divakar Yadav. (2014). "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis." IEEE 2014.

[3] Liu, Bing. (2010). Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, 2nd ed. Chapman and Hall: Florida.

[4] Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, Sweta Tiwari,"Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.8, No.4, pp.54-62, 2016. DOI: 10.5815/ijieeb.2016.04.07

[5] Meena Rambocas, João Gama, "Marketing Research: The Role of Sentiment Analysis", April 2013, ISSN: 0870-8541.

[6] Milan Gaonkar , Prof. Amit Patil."Sentiment Classification Using Product Reviews".

[7] Park, Do-Hyung, Lee, Jumin, and Han, Ingoo. (2007). "The Effect of On-Line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement." International Journal of Electronic Commerce, 11 (4): 125-148.

[8] 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.

[9] Rasika Wagh,Payal Punde."survey on sentiment analysis using twitter dataset." conference paper2018

[10] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," Proceedings of the Conference on Knowledge Capture (K-CAP), 2003.

[11]Customer Reviews Dataset, https://www.kaggle.com/vigneshwarsofficial/reviews