Review Sentylis

Bennett University
Computer Science Engineering
Yelda Jaswant Naidu (E19CSE071), Vamsi Jangala (E19CSE054)

INTRODUCTION

Our project is about building a machine learning model which analyzes review based on performance thereby generating a score and use sentimental analysis to check the polarity of situation wherein it can be a deciding factor. We also aim at building a model which generates a graphical representation of the data generated after cleaning the dataset, we tend to remove unpleasant or derogatory terms which affects the mental well-being of users or people who reads the reviews, last but not least we plan on automating this manual labor thus reducing the time which the user of the product spends on it and creating a safe and healthy platform for all

The introductory part is further divided into the following section

1. Problem statement.
2. Motivation for solving the problem.
3. Solution.
4. Background knowledge

Sentiment Analysis is a popular in machine learning to analyze the text and find the polarity like positive and negative. We use the sentiment analysis in machine learning model to predict the emotion of a people based on the text written.

While working on this project we mainly relied on the sentimental analysis factor which these days has been a go to approach for almost each and every platform not only its one of the most used common approach but also it has a wide range of usefulness like

• Emotion detection
• Aspect based analysis.
• Multilingual sentiment analysis and comes with many benefits like
• Scaling and sorting of data
• Real time analysis
• Time saving factor

MOTIVATION

For example, lets take a scenario wherein a visiting faculty gives a session on some topic and at the end of the day they want a review from the class now going through traditional methods of collection and reading each review the following things are likely to happen

• Time consuming and tiresome(labor).
• They may stumble upon some unpleasant words.
• A chance for missing out some genuine to the point review in the masses.
• A quick conclusion.

RELATED WORK

Sentimental analysis is used in a broad range of works relating from social media monitoring apps like twitter etc. to brand based review collection like Myntra/ajio therefore we attached examples of each under division\\

Social Media Monitoring:

Sentiment Analysis is widely used in the so many platforms to know the behavior of person based on their text. Sentiment Analysis is used in the social media monitoring to know the user feeling while using their product.

Let's us take about the twitter is a big social media network around the world. Twitter use the sentiment analysis and they allow companies to understand their users felling about their brand and its helps to know their competitors and they can focus on keep in demand on their product and introduce the new trends and keep on update to present trends needs. To perform the sentiment analysis on twitter has to follow steps. Initially twitter has to gather twitter data and they have clear idea about type of tweets they want. After that prepare the data and clean it includes the removal of irrelevant information and delete the duplicated data. So, we need to create the sentiment analysis machine learning model in this step we have to train the model to predict the emotions based on the data. We train the model in this phase to predict the positive and negative of statements from the following the data. After the sentiment analysis model training, we will analyze the twitter data for sentiment. They final step is visualize the results. Twitter sentiment analysis allows us to keep track about your product on social media it will detects the customer feelings before they escalate.

Brand monitoring:

Sentiment Analysis is used in brand monitoring. Brand monitoring is a basically know the what people think about their brand. Every company wants to know about their product in market it will beneficial to their brand. To get the information company tries to get it from the different number of channel and media. Company also searches the shopping sites like amazon reviews. With

the help of this they can able to improve their service to their product.

First Companies has to find the data comes from different ways like twitter and Facebook and shopping apps reviews etc. After that we need to filter the statements like positive or negative with the brand sentiment analysis algorithm. The major thing is analyzing the information it is important task is not only finding the negative and positive but need to know who, where, what the particular terms we need to keep in mind. We need to find which type of targeted people are giving the positive feedback and which type of targeted people are given negative feedback we need to analysis. So, with the help of this information the company can alerts and focus to improve the product.

The major step is Prevent the crisis. Let's us take an example your product has gave the negative feedback in shopping app we get notified by the sentiment analysis model. So, we to apologize the customer for the product and give the compensation voucher to prevent loss or not a mistake from company the company has to defend them self.

Customer feedback:

Sentiment Analysis is used in customer feedback. Basically, the customer feedback is the process of detecting the emotions of a customer interact with the products and services. In the customer feedback we mainly use the Natural language Processing and use the algorithm to detect the emotions of a customers. The algorithm main focus on the two parameters they are polarity and magnitude. Polarity is used for detect the positive and negative whereas magnitude is used for emotions exhibited by customer.
For performance of customer feedback, we need the data sets from different resources. Collect the data from the live chat. We need to collect the customer feedback every live chat. With the help of it we need to classify the feedback into respective terms. We need to collect the customer sentiment analysis from the social media. Social media platform such as twitter and Facebook we need to collect the data like "good"," bad"," hate", "awesome" such kind of positive or negative word. With the help of it we can collect data. To know our product, conduct the online surveys to get better in view of the customer side. Monitor the reviews and ratings in shopping apps like amazon, google play, Ajio. It is best way to predict the customer feelings and sentiment analysis.

PROPOSED METHODOLOGY

Data Collection
•The data is collected from the student feedback through the web portal. Every student has to login into the website with their specific credentials and give feedback and submit it.

Data Pre-Processing
•Wherein the given data which in raw or containing unwanted items get cleaned to form a new data set

1. Tokenization: Tokenization means splitting a sentence into a number of sub words like phrases, specific words are called tokens. In tokenization, it finds the punctuation and deleted from that sentence. Usually, tokenization means the splitting of a sentence into different small items. The tokenization process can be as follows: it converts the paragraph into a group of sentences and converts the sentences into words. Main thing in this removal of noise words from the sentence or paragraph.

2. Stop words removal: Stop words are must be filtered before the processing of data. We perform the stop word to get the proper and meaningful complete sentence. There are many stop words such as "and, the, of, before, why. What..". Stop word means it doesn't contain any meaningful information in the sentence. In any language, the stop words are familiar. For the making of NLP models, the stop words are not helpful in the experiment. The words like "pen, good, where" are called meaningful words.

3. Clustering: clustering is a process of creating the object into objects. The combination of the objects is formed into groups. While doing the clusters, the data is included in the supported data, and then it converts the label to a defined group. We find the advantage of the clustering is adaptable changes.

4. Classification: Classification is the process of making the data in a valuable and efficient way. When data is ideally classified, it is the model to predict with high accuracy. So best classification can decrease the errors and increase the efficient values.

5. Part of Speech Tagging: In speech tagging, the sentence or paragraph is compared with the speech part. For the part of speech, we have used the Stanford speech. In tagging, the data is split into some sentences and formed as part of the speech tag.

6.Feature Extraction: The process of feature extraction, every sentence is formed as some specific features. Sentiment score is useful for finding the polarity of a text. They are two types of sentiment words.

1. **Positive sentiment Words:** Sent WordNet defines the words as positive sentiment words and negative words. We can find the Positive sentiment word with

the help of Sent WordNet. The positive words like "good, impressive, great."

2. Negative sentiment Words: Sent WordNet define the word are in the positive sentiment words and negative words. We can find the Negative sentiment word with the help of Sent WordNet. The negative words like "not good, gawky. ghastly. grave. Greed."

Feature Reduction:

The high dimensions can show the impact on the performance of the classifier. When there is non-useful information in the data, there is a direct technique required to remove those sentences. So, with the help of that technology, we can differentiate the sentence into valuable words. We use the Gain ratio and Information Gain in the feature reduction technique.

Sentiment Dictionary:

The sentiment Dictionary consists of different kinds of lists and certain polarities like positive, negative, and neutral. So due to this kind of library make the user as user friendly. We have used different types of data have been used for the testing features of the record as follows.

1. First, the dictionary has to find the word that has priority as it is a strong or weak word.
2. the Second dictionary has to find the length of the word.
3. We need to represent the tokens for every word in the record.
4. We have to use the Part of speech for the words.
5. After the part of speech, the stemmed technique has to use. Every word is tally with the dictionary words and finds it behavior like "verb, adjective, noun."
6. The last step is finding the prior polarity. In this stage, every word is categorized into positive, negative, and neutral. So, we have to count the polarity with the help sentiment Dictionary.

Polarity Tagging:

The polarity tagging means tagging every word with respective of the polarity like positive, negative, neutral words with the help of sentiment Dictionary. In the tagging, the neutral words have been removed due to it doesn't contain any value it.

Let us take an example:
"This course gave me more knowledge, teacher help me to understand very well, basically the course is hard."

In this example, the text goes under the preprocessing stage and then every word is tagged with a polarity like positive, negative, neutral with the help of a sentiment dictionary. In the example, the hard work is tagged as the negative and the understanding, knowledge, and help are considered as positive tagged.

Sentiment words ---> hard knowledge understand help
Word polarity ---> negative positive

Word Frequency:
The word frequency means giving the value for every word. When the word occurs in only one time, then the frequency is one. Let us discuss this with the example.
Sentiment words ---> hard knowledge understand help
Word Frequency ---> 1 1 1 1
Word Attitude:
The Word Attitude means in this phase polarity converts into the numeric form. The numeric form helps us to do further processes.

$$\text{Word Attitude} = \begin{cases} 1, & \text{if the polarity is positive} \\ -1, & \text{if the polarity is negative} \end{cases}$$

Sentiment words ---> hard knowledge understand help
Word Attitude ---> -1 +1 +1 +1

Overall Attitude:

The Overall Attitude means the multiplication of both Word Attitude and Word Frequency.

Overall attitude = WA * WF

Sentiment words ---> hard knowledge understand help
Overall Attitude ---> -1 +1 +1 +1

Sentiment Score:

The sentiment Score means addiction of overall attitude score gives us the sentiment score. So, the sentiment score helps us to evaluate the teacher's performance. In sentiment score, we have to add the positive and negative words score. \

Sentiment Score = (+1) + (-1) + (+1) + (+1)
= 2.

\Section {Our approach}
Our main aim is to create a platform wherein we as a team are building a model which
can be summarized to a few points which are further elaborated, receive data (csv format)- real time data is collected in a csv file, clean data (remove derogatory words)- stop words etc. are removed in this process's which carry no weight to the overall data collection. certain sections like example:(package names for google reviews which carry no weight during the data analysis will be removed during data prepossessing. Then the final data which is now a clean data will be used for further evaluation to check which algorithm gives the best accuracy.

Naive Bayes

At first, we started with complement NB later on we improved and achieved the highest accuracy for the model using the Gaussian NB model.

Navie bayes algorithm is useful for the classification purpose. Navie bayes classifier has one feature different from other classifiers are when the data contain the particular feature it assumes that remaining it features are different from the set of class. Navie bayes are useful in dealing the large number of data classification. It gives the best performance model compare to other model classifiers.
Navie bayes is generally calculate the probability of each term and make it simplified with calculation. It takes the independent value targeted.

Navie Bayes model representation:

We can also call the representation are the probabilities of naive bayes.
They are two types of probabilities.
1.Class probability 2. conditional probability

In the class probability, it takes the probability of every training dataset. The second one is the conditional probability; it finds the probability of each input values.

Finding the class probability:

The class probability is found by dividing the number of instances to the total number of instances.

Prob(class=1) = count(class=1)/(count(class=0) + count(class=1))

Finding the conditional probabilities:

The conditional probabilities are found by dividing the frequency of each value of attribute to the total frequency of class.

Ex: Prob (weather = summer/school = holiday) = count (instance weather = summer and school = holiday)/count (instance with school = holiday)

$$P(x_i \mid C_k, x_1, \ldots, x_n) = P(x_i \mid C_k)$$

$$P(C_k \mid x_1, \ldots, x_n) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i \mid C_k)}{P(x_1, \ldots, x_n)}$$

$$P(C_k \mid x_1, \ldots, x_n) \propto P(C_k)\prod_{i=1}^{n} P(x_i \mid C_k)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_k P(C_k)\prod_{i=1}^{n} P(x_i \mid C_k)$$

$$\hat{y} = \arg\max_k \left(\ln P(C_k) + \sum_{i=1}^{n} \ln P(x_i \mid C_k)\right)$$

$$\hat{c} = \arg\max_{c \in C} P(c \mid d)$$

$$\hat{\theta}_{ki} = \frac{N_{ki} + \alpha}{N_k + \alpha n}$$

$$P(C_k \mid x_1, \ldots, x_n) = \frac{P(C_k)P(x_1, \ldots, x_n \mid C}{P(x_1, \ldots, x_n)}$$

Algorithm Modification and Accuracy:

```
#Accuracy using Naive Bayes Model
NB = MultinomialNB()
NB.fit(x, y)
y_pred = NB.predict(x_test)
print('\nNaive Bayes')
print('Accuracy Score: ',metrics.accuracy_score(
print('Confusion Matrix: ',metrics.confusion_mat
```

```
Naive Bayes
Accuracy Score: 85.65022421524664%
Confusion Matrix:
[[140   6]
 [ 26  51]]
```

Fig -NB ACCURACY

```
#Accuracy using KNN Model
KNN = KNeighborsClassifier(n_neighbors = 3)
KNN.fit(x, y)
y_pred = KNN.predict(x_test)
print('\nK Nearest Neighbors (NN = 3)')
print('Accuracy Score: ',metrics.accuracy_score(y_test,y_pred)*100,'%',sep='')
print('Confusion Matrix: ',metrics.confusion_matrix(y_test,y_pred), sep = '\n')
```

```
K Nearest Neighbors (NN = 3)
Accuracy Score: 38.11659192825112%
Confusion Matrix:
[[  9 137]
 [  1  76]]
```

Fig -KNN ACCURACY

```
#Accuracy using SVM Model
SVM = LinearSVC()
SVM.fit(x, y)
y_pred = SVM.predict(x_test)
print('\nSupport Vector Machine')
print('Accuracy Score: ',metrics.accuracy_score(y_test,y_pred)*100,'%',sep='')
print('Confusion Matrix: ',metrics.confusion_matrix(y_test,y_pred), sep = '\n')
```

```
Support Vector Machine
Accuracy Score: 81.16591928251121%
Confusion Matrix:
[[125  21]
 [ 21  56]]
```

Fig -SVM ACCURACY

```
#Accuracy using Logistic Regression Model
LR = LogisticRegression()
LR.fit(x, y)
y_pred = LR.predict(x_test)
print('\nLogistic Regression')
print('Accuracy Score: ',metrics.accuracy_score(y_test,y_pred
print('Confusion Matrix: ',metrics.confusion_matrix(y_test,y_
```

```
Logistic Regression
Accuracy Score: 82.51121076233184%
Confusion Matrix:
[[131  15]
 [ 24  53]]
```

Fig -LOGISTIC ACCURACY

After further changes in the algorithm, we have achieved the required accuracy using Support Vector Machines

```
# results
print("Results for SVC(kernel=linear)")
print("Training time: %fs; Prediction time: %fs" % (time_linear_train, time_linear_predict))
report = classification_report(testData['Label'], prediction_linear, output_dict=True)
print('positive: ', report['pos'])
print('negative: ', report['neg'])
```

```
Results for SVC(kernel=linear)
Training time: 8.941781s; Prediction time: 0.905810s
positive:  {'precision': 0.9191919191919192, 'recall': 0.91, 'f1-score': 0.9145728643216081, 'support': 100}
negative:  {'precision': 0.9108910891089109, 'recall': 0.92, 'f1-score': 0.9154228855721394, 'support': 100}
```

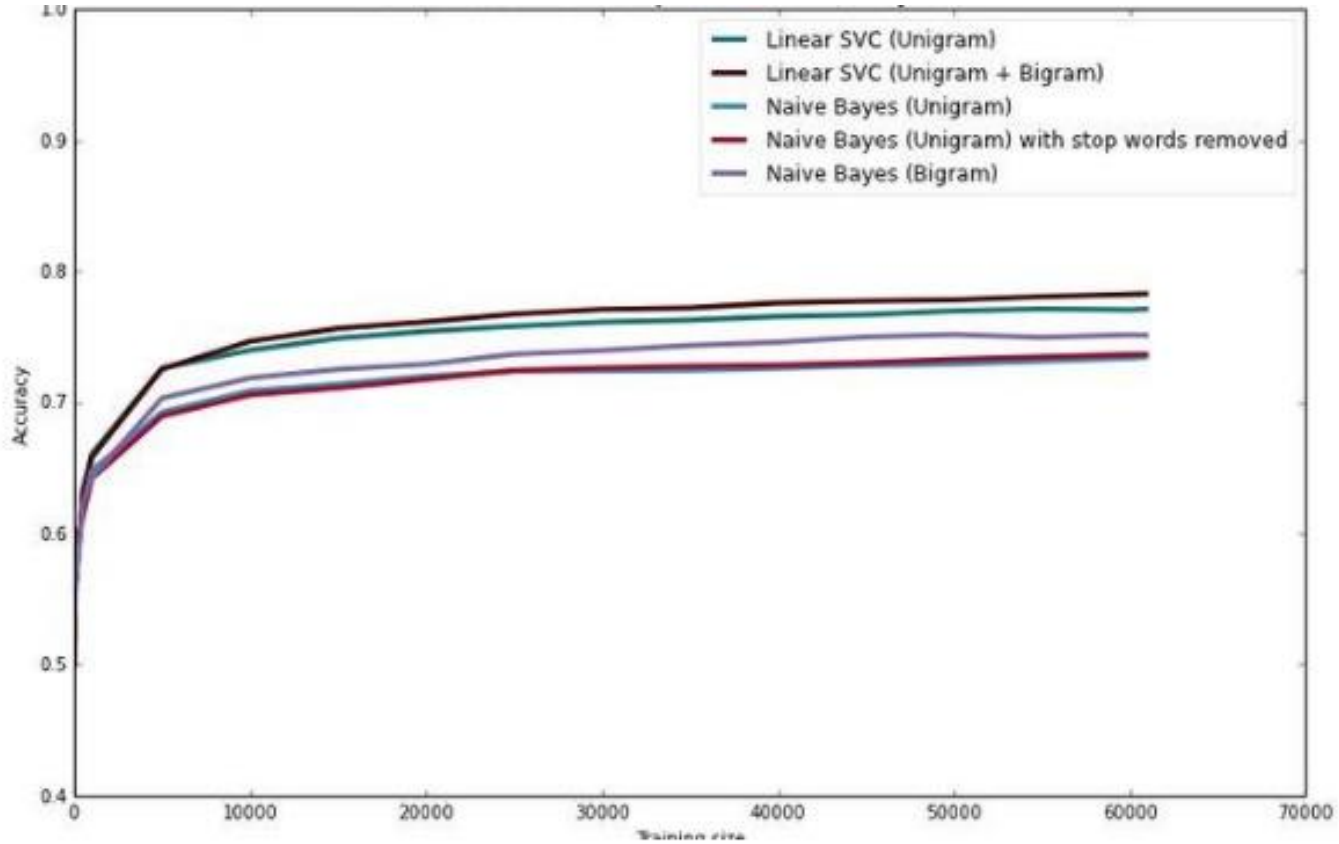
The F1 score , recall ,accuracy and Precision

Algorithm	Accuracy
Naïve Bayes (unigram)	74.56
Naïve Bayes (bigram)	76.44
Naïve Bayes (trigram)	75.41

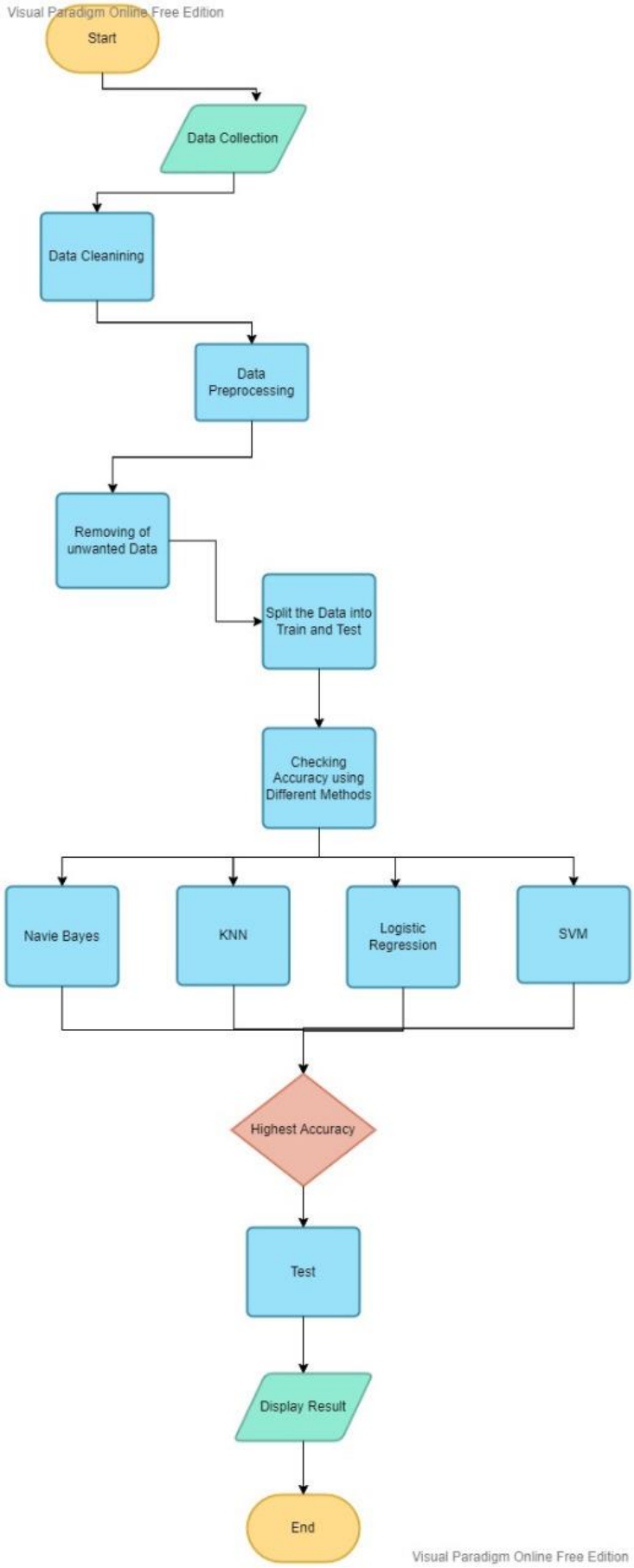
ACCURACY OF NB

SCORE	TEXT
2	You're awesome and I love you
-5	I hate and hate and hate. So angry. Die!
4	Impressed and amazed: you are peerless in your achievement of unparalleled mediocrity.

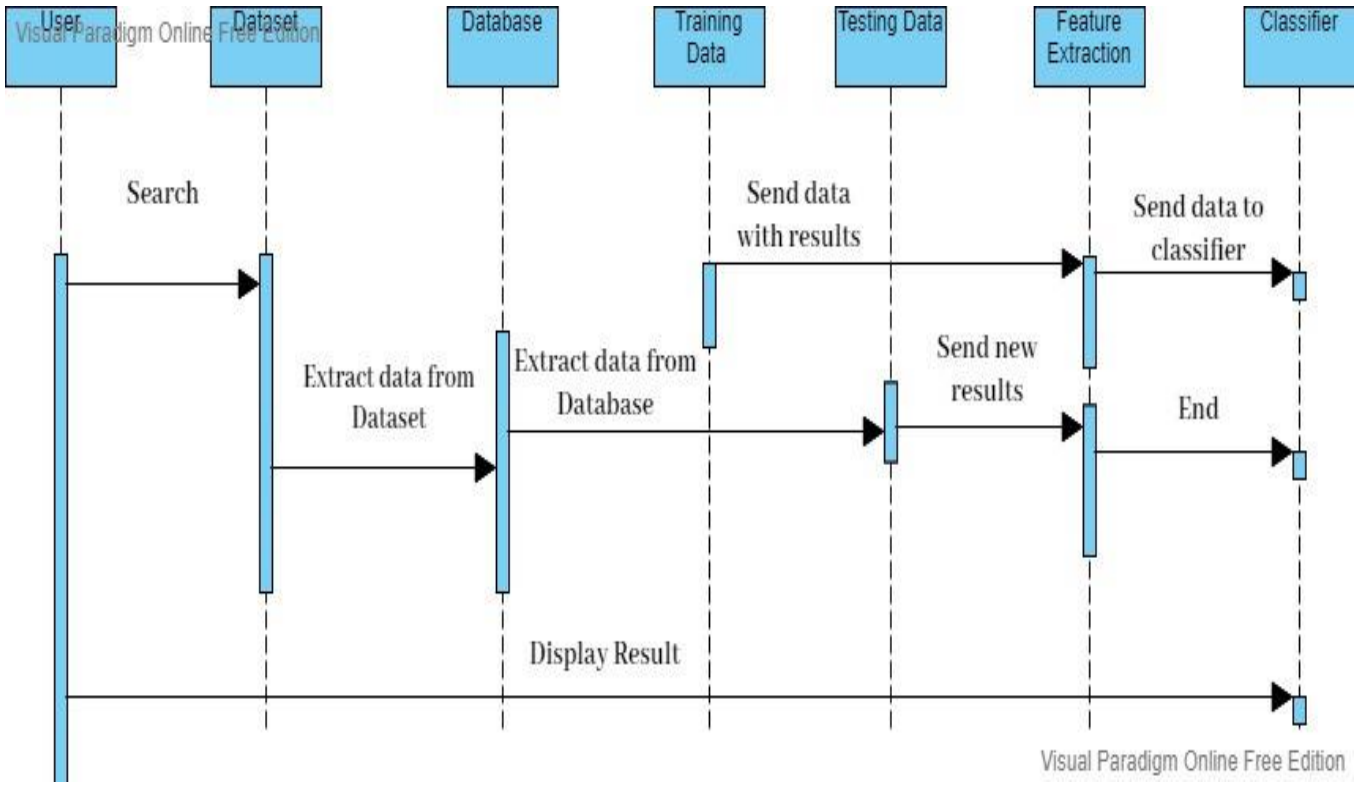
SCORE TO TEXT EXAMPLE



GRAPHICAL REPRESENTATION OVER ACCUARCY AND TRAINING DATA SET



FLOW DIAGRAM



::SEQUENCE DIAGRAM::

REFRENCES AND LINKS

[1] S. Rani and P. Kumar, "A Sentiment Analysis System to Improve Teaching and Learning," in Computer, vol. 50, no. 5, pp. 36-43, May 2017, Doi: 10.1109/MC.2017.133.\\

[2] K. L. S. Kumar, J. Desai and J. Majumdar, "Opinion mining and sentiment analysis on online customer review," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016, pp. 1-4, doi: 10.1109/ICCIC.2016.7919584. \\

[3]D. Zimbra, M. Ghiassi and S. Lee, "Brand-Related Twitter Sentiment Analysis Using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks," 2016 49th Hawaii International Conference on System Sciences (HICSS), 2016, pp. 1930-1938, Doi: 10.1109/HICSS.2016.244.\\

[4] M. H. Abd El-Jawad, R. Hodhod and Y. M. K. Omar, "Sentiment Analysis of Social Media Networks Using Machine Learning," 2018 14th International Computer Engineering Conference (ICENCO), 2018, pp. 174-176, doi: 10.1109/ICENCO.2018.8636124.\\

((Learning from Students' Perception on Professors Through Opinion Mining) [online] Available from:{https://link.springer.com/chapter/10.1007/978-3-030-61702-8_23 }
•https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

•https://www.ciodive.com/news/companies-using-sentiment-analysis-software-to-understand-employee-concerns/407357/

•https://www.wsj.com/articles/how-do-employees-really-feel-about-their-companies-1444788408

•https://www.researchgate.net/publication/51969319_Scikit-learn_Machine_Learning_in_Python

•https://scikit-learn.org/stable/modules/svm.html.

•https://towardsdatascience.com/naive-bayes-in-machine-learning-f49cc8f831b4

•https://www.tokenex.com/resource-center/what-is-tokenization#:~:text=Tokenization%20is%20the%20process%20of,the%20same%20length%20and%20format.

•https://monkeylearn.com/
•https://scikit-learn.org/stable/