EMOTION DETECTION

NATURAL LANGUAGE PROCESSING

PREPARED BY :- RAJ KUMAR

ABSTRACT

Sentiment analysis is a critical task in the field of Natural Language Processing (NLP) that aims to determine the emotional tone expressed in a piece of text, whether it's positive, negative, or neutral. In this report, we present a machine learning-based approach for text-based sentiment/emotion analysis. We know the importance of sentiment analysis in various applications, ranging from social media monitoring to customer feedback analysis. The primary focus of our work is to develop a robust and accurate sentiment/Emotion detector analysis model using state-of-the-art NLP techniques and machine learning algorithms. We share our methodology, experimental results, and insights into the challenges and opportunities within this domain.

In this project, we analyze different machine learning models to get the best model. The performance of all approaches are evaluated using accuracy score and precision and it attains an accuracy of 92% in Random forest and Bagging classifier.

Table of content

INTRODUCTION

Sentiment analysis, often referred to as opinion mining, is a subfield of Natural Language Processing (NLP) that involves the identification, extraction, and classification of emotional or subjective information from text data. It plays a crucial role in understanding how people express themselves online, in customer feedback, product reviews, social media, news articles, and a wide range of other textual sources. This information is invaluable for businesses, organizations, and researchers as it provides insights into public opinion, user satisfaction, and market trends.

Sentiment analysis has gained significant attention due to its applicability in diverse domains. For businesses, it helps in understanding customer sentiments and tailoring products and services accordingly. In the field of social media monitoring, sentiment analysis enables tracking public opinion and assessing the impact of campaigns or events. Media companies use sentiment analysis to gauge audience reactions to news articles or broadcasts. Moreover, it has applications in fields as diverse as healthcare, politics, and finance.

This report presents our endeavor to develop an emotion analysis model that effectively captures the nuances of sentiment/Emotion in textual data. We explore techniques in Natural Language Processing and machine learning to achieve this goal. Our approach involves data preprocessing, feature extraction, model selection, and performance evaluation. We discuss the challenges and limitations of sentiment analysis and provide recommendations for future research in this area. The goal of this work is to contribute to the growing body of knowledge in sentiment analysis and offer practical insights into building accurate and robust sentiment analysis systems.

METHODOLOGY

The working of the system starts with the collection of data. Then the required data is preprocessed using stemmer, stopword removal, textcase conversion etc. The data is then divided into two parts: training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy and precision of the system is obtained by testing the system using the testing data.

This system is implemented using the following modules.

1.)Collection of Dataset

2.)Data Preprocessing

3.)Model Building.

DATA COLLECTION

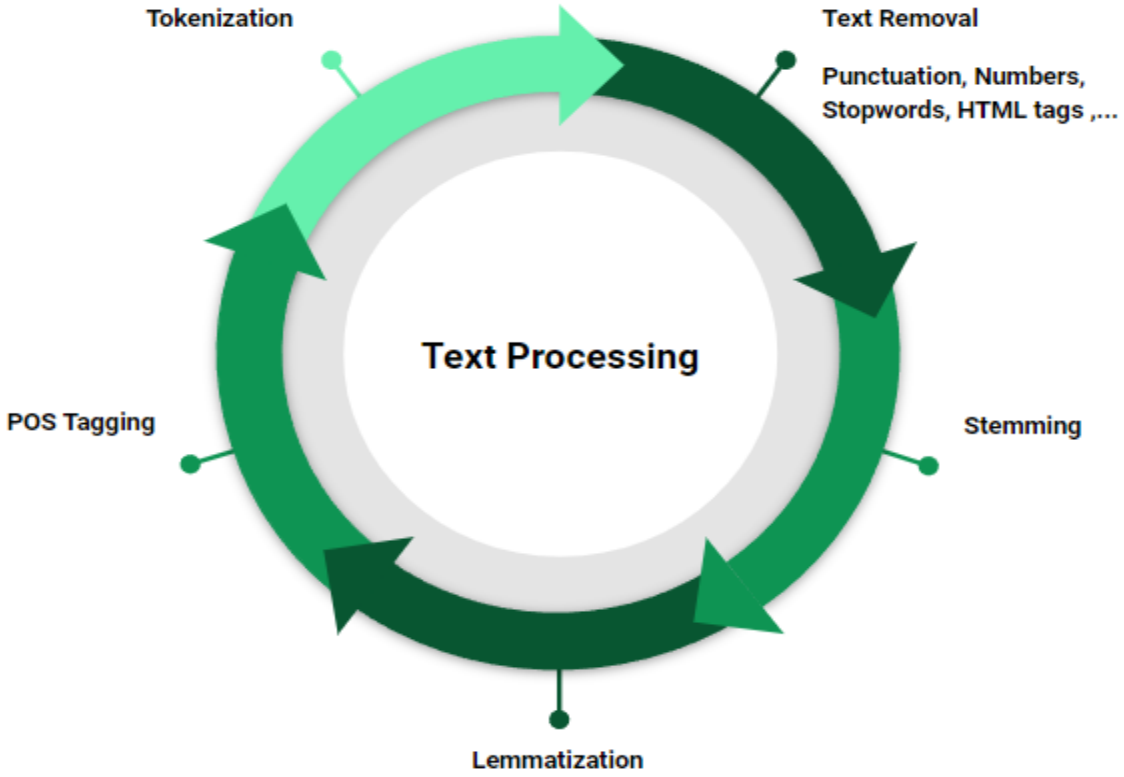
Initially, we collected a dataset for our Emotion detection system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning while testing data is used for evaluating the prediction model. For this project, 80% of the data is used for training and 20% of data is used for testing.

The dataset used for this project is taken from Kaggle https://www.kaggle.com/datasets/abdallahwagih/emotion-dataset. The dataset consists of 2 attributes.

PRE- PROCESSING OF DATA

Data preprocessing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes.

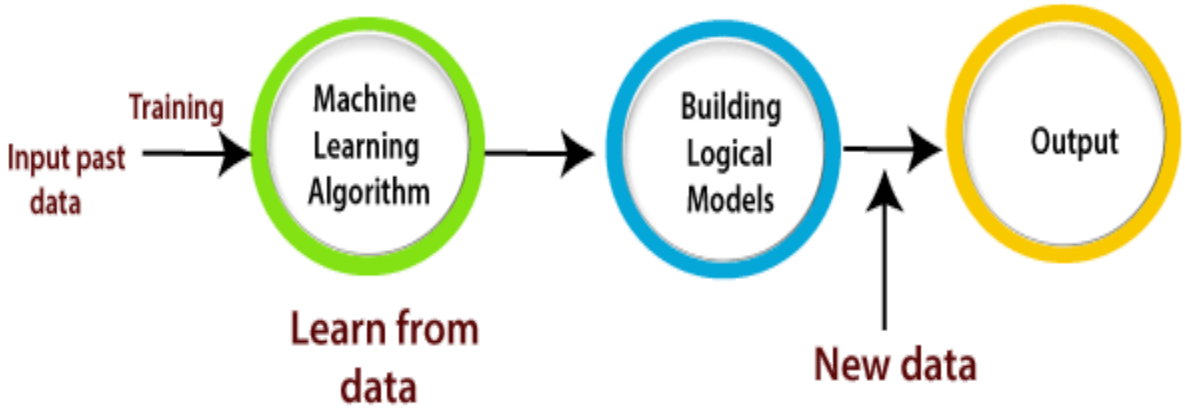
In pre-processing of data, we transform data into our required format. Data pre-processing has activities like tokenization, removing stop-words,stemming of words,vectorization of inputs etc. Preprocessing of data is required for improving the accuracy of the model.



MODEL BUILDING

Various machine learning algorithms like KNN,logistic regression,SVC, Decision Tree, ensemble models are used for Classification. Comparative analysis is performed among algorithms and the algorithm that gives the good outcome will be used for emotion detection.

Performance metrics are Accuracy score and Precision score are used to analyze models performance.

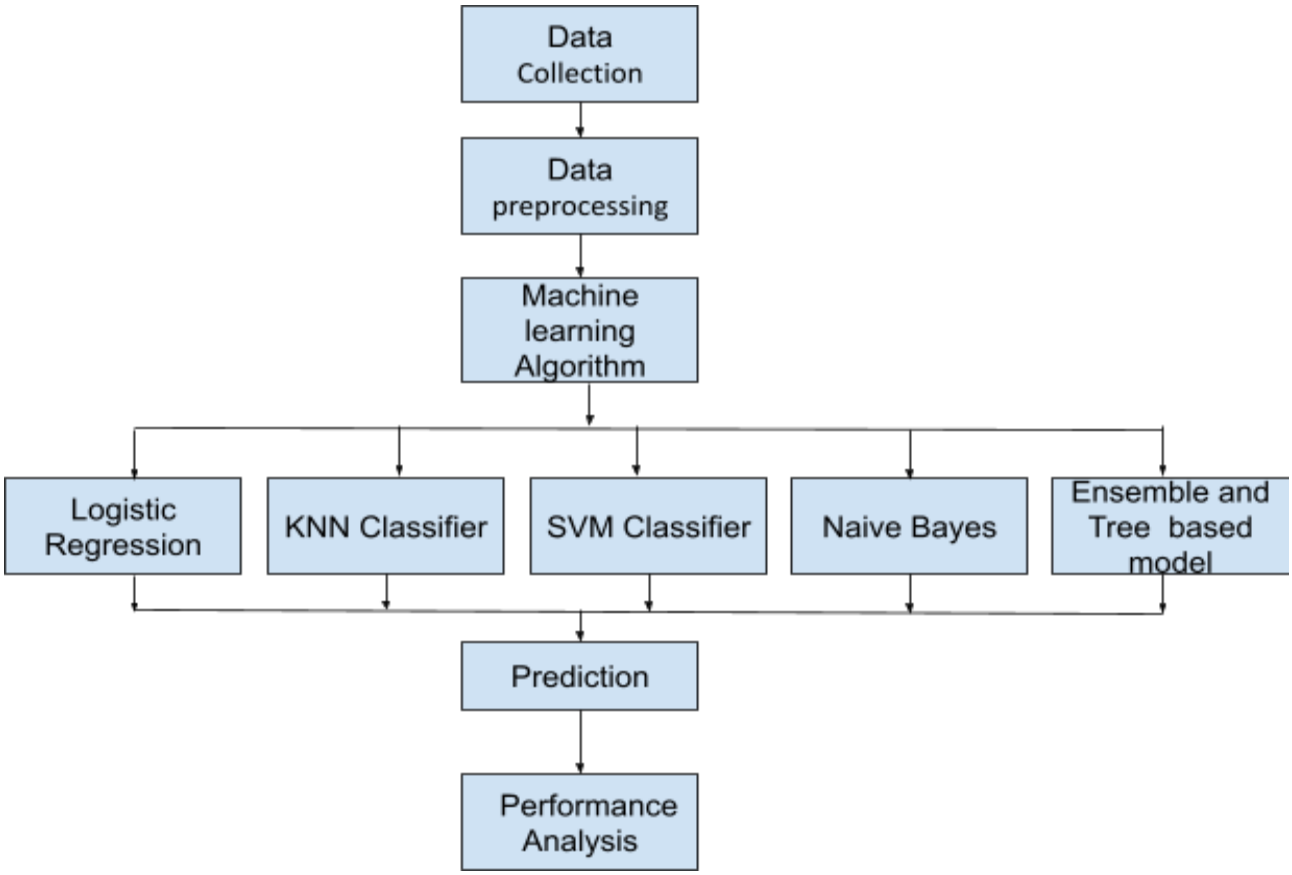


WORKING OF SYSTEM

SYSTEM ARCHITECTURE

The system architecture gives an overview of the working of the system.

The working of this system is described as follows: -
Dataset collection is collecting data which contains wine details. Data is further cleaned, made into the desired form. Different classification techniques as stated will be applied on preprocessed data to predict the performance of models.



DATASET DESCRIPTION

Dataset contains two features:-

- Comment:- Messages, based on these emotion will detect
- Emotion:- Describing emotion of messages

From the dataset we can see we have 5937 rows and 2 features.

LIBRARIES USED

Python libraries make it easy for us to handle the data and perform typical and complex tasks with a single line of code.

Libraries used in this case study are as follow:-

- **Pandas**– Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

- **NumPy**–NumPy stands for Numerical Python. NumPy is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra, fourier transform, and matrices.

- **Matplotlib**– Matplotlib is a low level graph plotting library in python that serves as a visualization utility.

- **Seaborn**– Seaborn is a Python visualization library based on matplotlib. It

provides a high-level interface for drawing attractive statistical graphics.

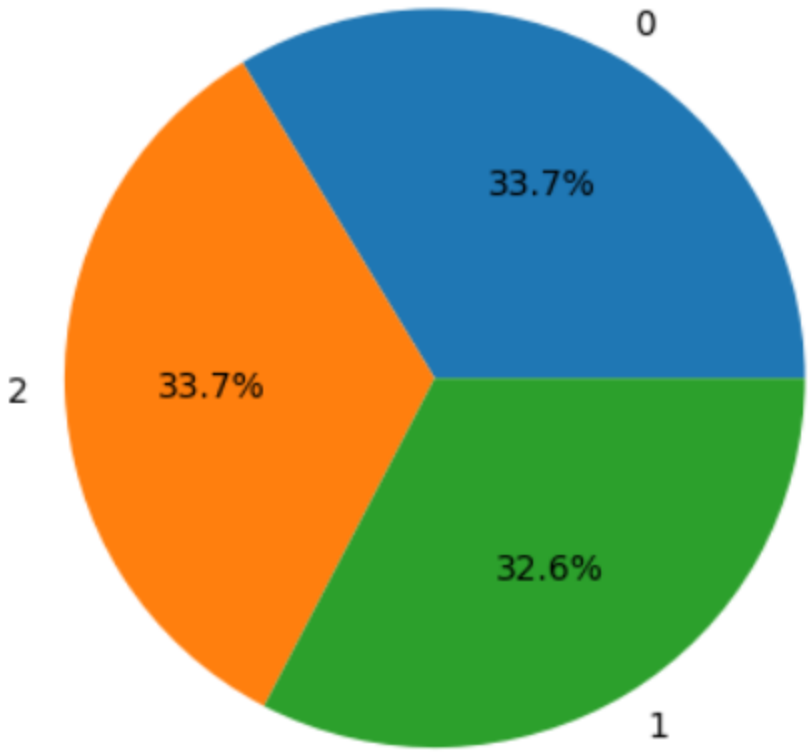
- **Sklearn**– This module contains multiple libraries having pre-implemented functions to perform tasks from data preprocessing to model development and evaluation.
- **Natural Language Toolkit (NLTK)-** NLTK is a popular Python library for working with human language data. NLTK includes various text processing libraries for classification, tokenization, stemming, tagging, parsing, and more.

DATA PREPROCESSING & EDA

Dataset contains two features one is comment and second is Emotion, emotion is our target in this case. Classes of Emotion feature are encoded in this form:- if Emotion=0 means "anger",if Emotion=1 means "fear",if Emotion=2 means "joy".

- Pie chart distribution of Emotion

Data is balanced, as data is distributed in almost the same percentage.



PREPROCESSING

In data preprocessing we first convert texts into lower cases as python is case sensitive language . It will treat lower case text and uppercase texts as different.

Tokenization:- after converting into lower cases we tokenize sentences, tokenize break sentences into words.

After tokenization we remove stop words from sentences and then perform stemming to get the root words of words in sentences.

After perform all preprocessing steps as above, final data is as follow:-

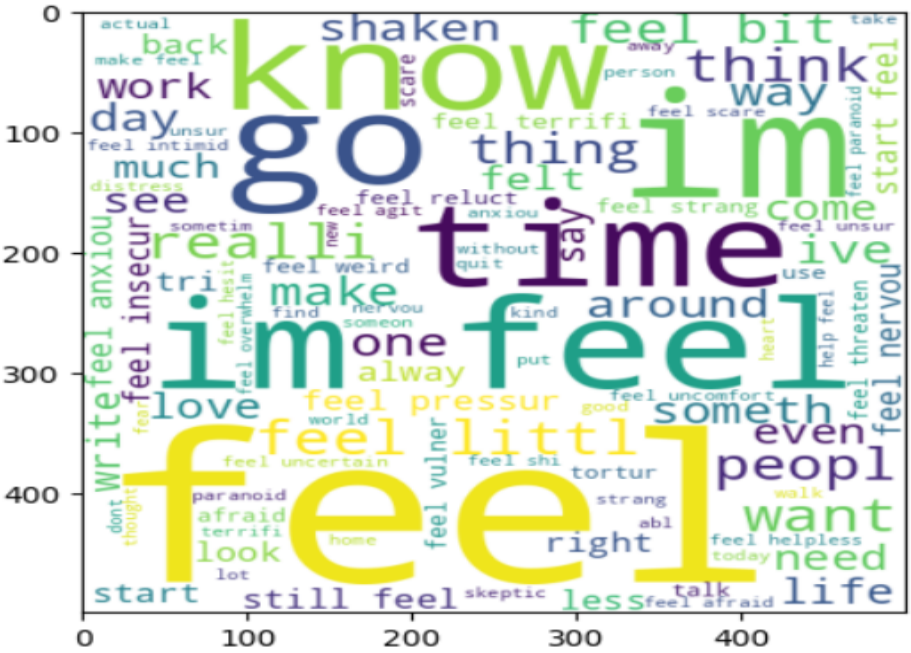
	Comment	Emotion	final_data
0	i seriously hate one subject to death but now ...	1	serious hate one subject death feel reluct drop
1	im so full of life i feel appalled	0	im full life feel appal
2	i sit here to write i start to dig out my feel...	1	sit write start dig feel think afraid accept p...
3	ive been really angry with r and i feel like a...	2	ive realli angri r feel like idiot trust first...
4	i feel suspicious if there is no one outside l...	1	feel suspici one outsid like raptur happen someth

WORD CLOUD

A word cloud is a visual representation of text data, where the size of each word is proportional to its frequency or importance within the text. It's a popular way to display the most frequently occurring words in a corpus of text, making it easy to identify and visualize key terms or topics.

- Word cloud for Emotion “Anger”

- Word cloud for Emotion “Fear”



- Word cloud for Emotion “Joy”

MODEL BUILDING AND EVALUATION

We will use this Emotion detection dataset for training and testing in the ratio of 80-20. The training dataset is used for model learning while testing data is used for evaluating the prediction model.
We will evaluate model performance with different algorithms.

EVALUATION

To evaluate our model we will use accuracy score and precision score as performance metrics.

LOGISTIC REGRESSION MODEL

ACCURACY SCORE

- accuracy score on train data is 0.9919983154348284
- accuracy score on test data is 0.9225589225589226

PRECISION SCORE

- precision score on train data is 0.991995791449933
- precision score on test data is 0.9226309573657144

SUPPORT VECTOR CLASSIFIER

ACCURACY SCORE

- accuracy score on train data is 0.9894714676774058
- accuracy score on test data is 0.9057239057239057

PRECISION SCORE

- precision score on train data is 0.9896740968685487
- precision score on test data is 0.9082367038796687

DECISION TREE CLASSIFIER

ACCURACY SCORE

- accuracy score on train data is 0.9991577174141925
- accuracy score on test data is 0.9284511784511784

PRECISION SCORE

- precision score on train data is 0.9991744066047472
- precision score on test data is 0.9300199258661005

RANDOM FOREST CLASSIFIER

ACCURACY SCORE

- accuracy score on train data is 0.9991577174141925
- accuracy score on test data is 0.9217171717171717

PRECISION SCORE

- precision score on train data is 0.999162881931753
- precision score on test data is 0.9236440128176344

NAIVE BAYES CLASSIFIER

ACCURACY SCORE

- accuracy score on train data is 0.9991577174141925
- accuracy score on test data is 0.9284511784511784

PRECISION SCORE

- precision score on train data is 0.9991744066047472
- precision score on test data is 0.9300199258661005

XGBOOST CLASSIFIER

ACCURACY SCORE

- accuracy score on train data is 0.9633607075173721
- accuracy score on test data is 0.9158249158249159

PRECISION SCORE

- precision score on train data is 0.9645737255935662
- precision score on test data is 0.9180274325919316

ADABOOST CLASSIFIER

ACCURACY SCORE

- accuracy score on train data is 0.6081280269530428
- accuracy score on test data is 0.5858585858585859

PRECISION SCORE

- precision score on train data is 0.6443241454560066
- precision score on test data is 0.6011900198726416

GRADIENT BOOST CLASSIFIER

ACCURACY SCORE

- accuracy score on train data is 0.9359865234786271
- accuracy score on test data is 0.8880471380471381

PRECISION SCORE

- precision score on train data is 0.9388411566926195
- precision score on test data is 0.8909753318629837

BAGGING CLASSIFIER

ACCURACY SCORE

- accuracy score on train data is 0.990945462202569
- accuracy score on test data is 0.9301346801346801

PRECISION SCORE

- precision score on train data is 0.9910534220408834
- precision score on test data is 0.9322192870966934

KNN CLASSIFIER

ACCURACY SCORE

- accuracy score on train data is 0.8026953042745841
- accuracy score on test data is 0.6632996632996633

PRECISION SCORE

- precision score on train data is 0.8239514197278949
- precision score on test data is 0.7032963022958322

CONCLUSION

We started with the data preprocessing where we checked for missing data. Afterward, we perform preprocessing using the NLP library on the dataset to extract all the words from sentences. Then after preprocessing started training on different machine learning models, and for each model, we conclude with the resultant Accuracy score and Precision score.

Based on prediction we conclude -

- Decision Tree, Random Forest, and Bagging models stand out with high precision and accuracy scores, making them strong candidates for reliable emotion detection.
- Logistic Regression and Support Vector Classifiers also perform well and provide consistency in precision and accuracy.
- Models like AdaBoost and KNN exhibit comparatively lower precision and accuracy; they may not be the best choices for precision-critical emotion detection tasks.

FUTURE SCOPE OF IMPROVEMENTS

As we can see the prediction on models , Some models are performing very well with training data but not able to retain the same performance with test/unseen data.While training, loss also occurs in some models. Hence, The model is overfitted.

On the other hand data is very less, we can train our model by adding more sentences, if more data will be used for the learning process, the model will predict well.

In this case we can try to improve model performance by tuning.

REFERENCES

https://scikit-learn.org/stable/supervised_learning.html

https://www.deeplearning.ai/resources/natural-language-processing/

https://www.geeksforgeeks.org/generating-word-cloud-python/

https://matplotlib.org/stable/plot_types/index.html