**SENTIMENT ANALYSIS**

**Project ID:** 29041

*B.Tech. Project Report*

*submitted for fulfillment of*

*the requirements for the*

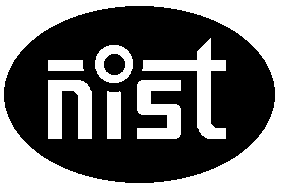
*Degree of Bachelor of Technology*

*Under Biju Pattnaik University of Technology*

*Submitted By*

**Rajesh Kumar Pradhan Sangram Kumar Mistri**

**Roll No : 202010292 Roll No: 202010281**



*2023 – 2024*

*Under the guidance of*

**Mr. Debasish Padhy**

**NIST INSTITUTE OF SCIENCE & TECHNOLOGY (Autonomous)**

**Palur Hills, Berhampur, Odisha – 761008, India**

# ACKNOWLEDGEMENT

We would like to take this opportunity to thank all those individuals whose invaluable contribution in a direct or indirect manner has gone into the making of this project a tremendous learning experience for me.

It is our proud privilege to epitomize our deepest sense of gratitude and indebtedness to our faculty advisor, **Mr. Debasish Padhy** for her valuable guidance, keen and sustained interest, intuitive ideas and persistent endeavor. Her guidance and inspirations enabled us to complete our report work successfully.

We give our sincere thanks to **Ms. Ch. Shree Kumar, Project Co-Ordinator**, for giving us the opportunity and motivating us to complete the project within stipulated period of time and providing a helping environment.

We acknowledge with immense pleasure the sustained interest, encouraging attitude and constant inspiration rendered by **Prof. (Dr.)** **Sukant K. Mohapatra** (Chairman), **Prof. (Dr.) Priyadarshi Tripathy** (Principal), **Prof. (Dr.)** **Brojo Kishore Mishra** (HOD , School of Computer Science) N.I.S.T. Their continued drive for better quality in everything that happens at N.I.S.T. and selfless inspiration has always helped us to move ahead.

Sangram Kumar Mistri(CSE202010281)

Rajesh Kumar Pradhan(CSE202010292)

# ABSTRACT

Every customer facing industry (retail, telecom, finance, etc.) is interested in identifying their customers’ sentiment, whether they think positive or negative about them. Python sentiment analysis is a methodology for analyzing a piece of text to discover the sentiment hidden within it. It accomplishes this by combining machine learning and natural language processing (NLP). Sentiment analysis allows you to examine the feelings expressed in a piece of text. IMDB Movie reviews is one of most popular social networking site where people are expressing their views, opinion and emotions liberally. These Reviews are recorded and analyzed to mine emotions of people related to it.

To mine emotions and polarity in Reviews, text mining techniques are used. Approximately 5000 Reviews are recoded and pre-processed to create a dataset of frequently appearing words. Every customer facing industry (retail, telecom, finance, etc.) is interested in identifying their customers’ sentiment, whether they think positive or negative about them. Sentiment analysis is a methodology for analyzing a piece of text to discover the sentiment hidden within it. It accomplishes this by combining machine learning and natural language processing (NLP). Sentiment analysis allows you to examine the feelings expressed in a piece of text.

Table of Contents

ACKNOWLEDGEMENT…………………………………………………………………….1

ABSTRACT……………………………………………………………………………….......2

TABLE OF CONTENTS……………………………………………………………………...3

1. PROBLEM STATEMENT………………………………………………………....4

2 OBJECTIVES ………………………………………………………………………5

3.INTRODUCTION ………………………………………………………………….6

4. METHODOLOGY

4.1-Data Collection……………………………………………………………….…7

4.2-Pre-Processing of Data…………………………………………..........................8

4.3-Feature Extraction………………………………………………………………..9

4.4-Model Implementation…………………………………………………………..10

5. APPLICATIONS …………………………………………………………………………11

6. REFERENCES……………………………………………………………………………12

# 1. PROBLEM STATEMENT

* The challenge lies in developing an effective sentiment analysis system that can accurately interpret and classify the emotional tone of diverse texts, such as social media posts, customer reviews, and news articles.
* We want to teach computers to understand these confusing situations better also want to make sure that computers don't make unfair or wrong judgments about people's feelings. We want them to be good at this job and not hurt anyone's feelings or make mistakes
* By addressing these challenges, we aim to provide a valuable tool for businesses to measure public opinion, enhance customer satisfaction, and make informed decisions based on the sentiment conveyed in textual data.

# 2. OBJECTIVE

The core objective of this project is to classify the sentiment behind the texts-

Python sentiment analysis is a methodology for analyzing a piece of text to discover the sentiment hidden within it. It accomplishes this by combining machine learning and natural language processing (NLP). Sentiment analysis allows you to examine the feelings expressed in a piece of text.

The aim while performing sentiment analysis on IMDB movie review is basically to classify the review in different sentiment classes accurately. In this field of research, various approaches have evolved, which propose methods to train a model and then test  
it to check its efficiency. Performing sentiment analysis is challenging on Review, as we mentioned earlier. Here we define the reasons for this:

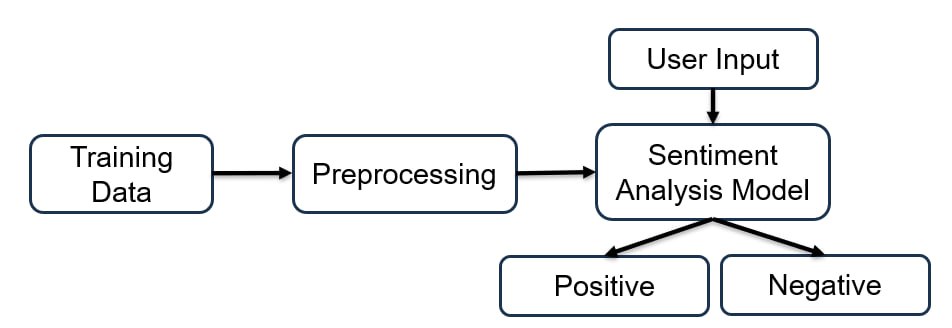


Fig-2

# 3.INTRODUCTION

Sentiment analysis on IMDB movie reviews involves using natural language processing (NLP) techniques to analyze and categorize the sentiments expressed in user reviews on the IMDB platform.

IMDB (Internet Movie Database) is a popular online database that provides information about films, television programs, and video games. Users on IMDB often submit reviews to share their opinions and experiences with movies, making it a rich source of textual data for sentiment analysis.

Sentiment analysis on IMDB reviews aims to automatically determine the sentiments expressed by users toward a particular movie. The sentiments are typically categorized as positive, negative.

This analysis can provide valuable insights into the reception of a movie among the audience and help filmmakers, producers, and viewers make informed decisions.

**What is sentiment analysis?**Sentiment analysis is a technique through which you can analyze a piece of text to determine the sentiment behind it. It combines machine learning and natural language processing (NLP) to achieve this.

Using basic Sentiment analysis, a program can understand whether the sentiment behind a piece of text is positive, negative, or neutral.

It is a powerful technique in Artificial intelligence that has important business applications.

For example, you can use sentiment analysis to analyze customer feedback. After collecting that feedback through various mediums like Movie review and Facebook, you can run sentiment analysis algorithms on those text snippets to understand your customers' attitude towards your product.

**Entertainment Industry:**

Film Production and Marketing: Filmmakers and producers can analyze IMDB reviews to understand audience reactions to their movies. Positive sentiments can guide marketing efforts, while negative sentiments can highlight areas for improvement.

**Business and Brand Management:**

Product Reviews: Companies use sentiment analysis to analyze customer reviews of their products or services. Positive feedback can be leveraged for marketing, while negative feedback can prompt improvements.

**E-commerce:**

Product Recommendations: Sentiment analysis contributes to personalized product recommendations by understanding customer preferences based on sentiment expressed in reviews.

**Healthcare:**

Patient Feedback: Sentiment analysis is used in healthcare to analyze patient feedback and reviews about hospitals, clinics, and healthcare providers, helping improve patient satisfaction.

Drug and Treatment Evaluation: Analyzing sentiments in medical literature and research papers can provide insights into the effectiveness and reception of different drugs and treatments.

**Education**:

Student Feedback: Educational institutions use sentiment analysis to analyze feedback from students, helping improve the quality of education and services.

Course Evaluation: Sentiment analysis is applied to evaluate sentiments expressed in course reviews, aiding in curriculum development

# 4. METHODOLOGY

In order to perform sentiment analysis, we are required to collect data from the desired source (IMDB movie review). This data undergoes various steps of pre-processing which makes it more machine sensible than its previous form.

The basic steps for performing sentiment analysis includes data collection, pre-processing of data, feature extraction, selecting baseline features, sentiment detection and performing classification.

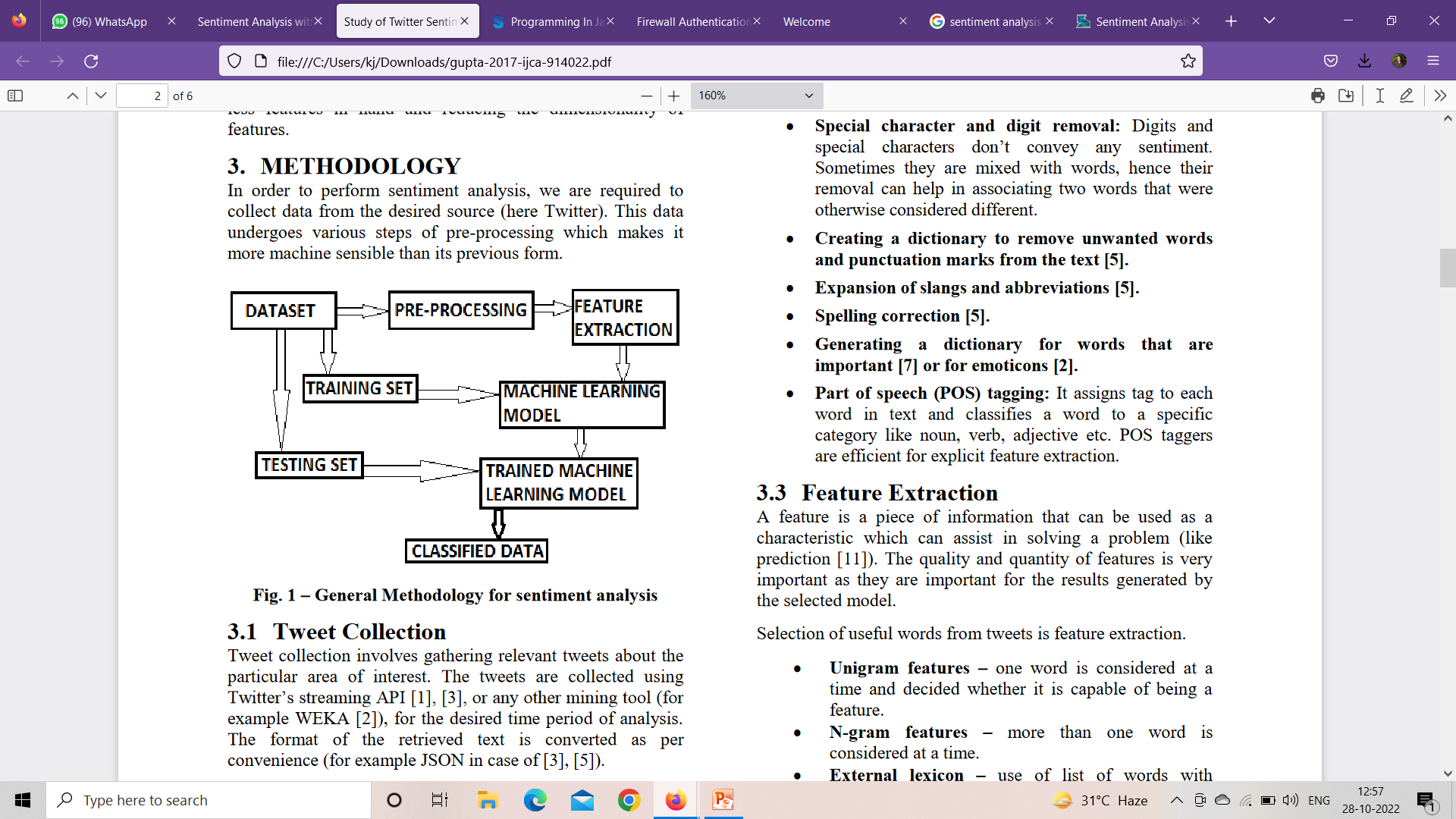


Fig - 3

# 4.1-Data Collection

Data collection involves gathering relevant IMDB movie review about the particular area of interest. The reviews are collected using IMDB Data set from Kaggle. The dataset collected is imperative for the efficiency of the model. The division of dataset into training and testing sets is also a deciding factor for the efficiency of the model. The training set is the main aspect upon which the results depends.

# 4.2-Pre-processing of Data

The preprocessing of the data is a very important step as it decides the efficiency of the other steps down in line. It involves syntactical correction of the Reviews as desired. The steps involved should aim for making the data more machine readable in order to reduce ambiguity in feature extraction. Below are a few steps used for pre-processing of Reviews .

• Converting upper case to lower case: In case we are using case sensitive analysis, we might take two occurrences of same words as different due to their sentence case. It is important for an effective analysis not to provide such misgivings to the model.

• Stop word removal: Stop words that do not affect the meaning of the reviews are removed (for example and, or, still etc.).

• Stemming: Replacing words with their roots, reducing different types of words with similar meanings. This helps in reducing the dimensionality of the feature set.

• Special character : Digits don’t convey any sentiment. Sometimes they are mixed with words; hence their removal can help in associating two words that were otherwise considered different.

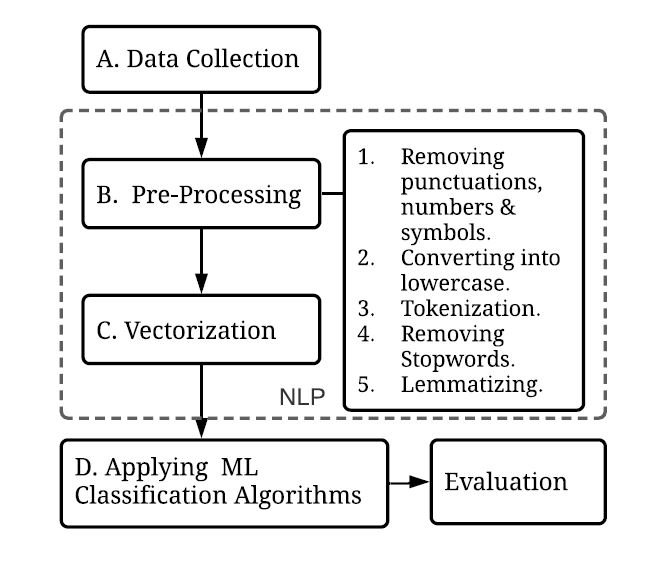


Fig-4

# 4.3-Feature Extraction

This line creates an instance of the TfidfVectorizer class with the specified parameters. max\_features is set to 5000, which means the vectorizer will only consider the top 5000 most frequently occurring words in the corpus.

When you use this vectorizer to transform a collection of text documents, it will convert the raw text into a numerical representation based on the TF-IDF values of the words. TF-IDF is a numerical statistic that reflects the importance of a word in a document relative to a collection of documents (corpus).

**Need of feature extraction techniques:**

Machine Learning algorithms learn from a pre-defined set of features from the training data to produce output for the test data. But the main problem in working with language processing is that machine learning algorithms cannot work on the raw text directly. So, we need some feature extraction techniques to convert text into a matrix(or vector) of features. Some of the most popular methods of feature extraction are :

1. Bag-of-Words
2. TF-IDF

**Bag of Words:** Bag-of-Words is one of the most fundamental methods to transform tokens into a set of features. The Bag-of-Words model is used in document classification, where each word is used as a feature for training the classifier. For example, in a task of review based sentiment analysis, the presence of words like *‘fabulous’, ‘excellent’* indicates a positive review, while words like *‘annoying’, ‘poor’* point to a negative review . There are 3 steps while creating a Bag-of-Words model:

1. The first step is **text-pre-processing** which involves:
   1. Converting the entire text into lower case characters.
   2. Removing all punctuations and unnecessary symbols.
2. The second step is to **create a vocabulary** of all unique words from the corpus.

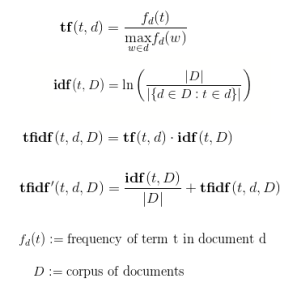
**TF-IDF Vectorizer:** TF-IDF stands for *term frequency-inverse document frequency*. It highlights a specific issue which might not be too frequent in our corpus but holds great importance. The TF–IFD value increases proportionally to the number of times a word appears in the document and decreases with the number of documents in the corpus that contain the word. It is composed of 2 sub-parts, which are:

1. Term Frequency (TF)
2. Inverse Document Frequency (IDF)

**Term Frequency (TF):** Term frequency specifies how frequently a term appears in the entire document. It can be thought of as the probability of finding a word within the document. It calculates the number of times a word occurs in a review , with respect to the total number of words in the review.

**Inverse Document Frequency (IDF):** The inverse document frequency is a measure of whether a term is rare or frequent across the documents in the entire corpus. It highlights those words which occur in very few documents across the corpus, or in simple language, the words that are rare have high IDF score.

**Term Frequency-Inverse Document Frequency (TF-IDF)** TF-IDF is the product of TF and IDF. It is formulated as:



A high TF-IDF score is obtained by a term that has a high frequency in a document, and low document frequency in the corpus. For a word that appears in almost all documents, the IDF value approaches 0, making the tf-idf also come closer to 0.TF-IDF value is high when both IDF and TF values are high i.e. the word is rare in the whole document but frequent in a document.

Fig - 5

# 9.APPLICATIONS

Movie Recommendation Systems:

Sentiment analysis can be integrated into movie recommendation systems to enhance the accuracy of recommendations. By analyzing user reviews, the system can better understand the sentiments associated with different movies and suggest films that align with a user's preferences

.

Marketing and Promotion:

Sentiment analysis can assist in marketing strategies. Positive sentiments in reviews can be used in promotional materials, while negative sentiments can highlight areas for improvement or adjustments in marketing campaigns.

Audience Engagement:

Film studios and marketing teams can use sentiment analysis to engage with the audience more effectively. By understanding how viewers feel about specific aspects of a movie, such as characters, plot, or cinematography, they can tailor their engagement strategies accordingly.

**Business and Brand Management:**

Product Reviews: Companies use sentiment analysis to analyze customer reviews of their products or services. Positive feedback can be leveraged for marketing, while negative feedback can prompt improvements.

# 11.REFERECES

[1] Mondher Bouazizi and Tomoaki Ohtsuki, “Sentiment Analysis: from Binary to Multi-Class Classification”, IEEE ICC 2016 SAC Social Networking, ISBN 978-1- 4799-6664-6.

[2] Halima Banu S and S Chitrakala, “Trending Topic Analysis Using Novel Sub Topic Detection Model”, (IEEE) ISBN- 978-1-4673-9745-2, 2016.

[3] Shi Yuan, Junjie Wu, Lihong Wang and Qing Wang, “A Hybrid Method for Multi-class Sentiment Analysis of Micro-blogs”, ISBN- 978-1-5090-2842-9, 2016.

[4] Neethu M S and Rajasree R, “Sentiment Analysis in Twitter using Machine Learning Techniques”, IEEE – 31661, 4th ICCCNT 2013.

[5] W. Y. Chong, B. Selvaratnam and L. -K. Soon, "Natural Language Processing for Sentiment Analysis: An Exploratory Analysis on Tweets," 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, Kota Kinabalu, Malaysia, 2019,

[6]Jaspreet Singh, Gurvinder Singh &Rajinder Singh “Optimization of sentiment analysis using machine learning classifiers” published on 11th December 2017

[7]Poria S, Cambria E, Gelbukh A, Bisio F, Hussain A (2015) Sentiment data flow analysis by means of dynamic linguistic patterns. IEEE Comput Intell Mag 10(4):26–36.

[8]Cambria E, White B (2014) Jumping NLP curves: a review of natural language processing research. IEEE Comput Intell Mag 9(2):48–57