Sentiment Analysis

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology/Master of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Edara Vara Siddha Vignesh Reg No: AP20110010058

Under the Guidance of

prof. Radha Guha

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[December, 2022]

Certificate

Date: 15-Dec-22

This is to certify that the work present in this Project entitled "**SENTIMENT ANALYSIS**" has been carried out by "**Edara Vara Siddha Vignesh [Reg No: AP20110010058]** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in **School of Engineering and Sciences**.

Supervisor

(Signature)

Prof. Radha Guha

professor,

Department of Computer Science and Engineering.

Acknowledgements

First and foremost we would like to thank my mentor (or) supervisor prof. Radha Guha mam for spending her time as well as effort in our research project her suggestions and some valuable information was really helpful for us to do our research project.

Last but not least we would like to thank to the college management for encouraging us all the time to do a research project and manage our workload I want to express my gratitude to everyone who helped us stay motivated to finish our research on time.

Table of Contents

Abstract

Text is the biggest body of information that humans have acquired over many years. The significance of this knowledge will be much enhanced if other insights are pounded into it. Sentiment Analysis (SA), which makes use of Natural Language Processing (NLP), provides a traditional machine learning (ML) approach to this problem. In the suggested experiment, we utilised SA on the dataset of IMDb movie reviews from Kaggle's Bag of Words to demonstrate how insightful information may be gleaned from a huge amount of textual data collected from the internet.

Four well-known machine learning (ML) methods help us get these insights: Nave Bayes (NB), Logistic Regression (LR). The Area under the Curve, Accuracy, Precision, Recall, Accuracy, and Confusion Matrix were utilised as the six assessment criteria to compare the performance of these two algorithms.

Abbreviations

SA	Sentiment Analysis
NLP	Natural Language Processing
IMBD	Internet Movie Database
AUC	Area Under Curve
DTM	Document Term Matrix
NB	Navie Bayes
ML	Machine Learning
LR	Logistic Regression

List of Tables

List of Figures

List of Equations

1. Introduction

What is sentiment analysis

It is only technique that recognizes the emotional content of a body of text using statistics, natural language processing and machine learning techniques like naïve bayes and logistic regression. Positive, negative, or neutral are all acceptable.

1.1purpose of sentiment analysis

Sentiment analysis (SA) seeks to identify the emotion underlying the data. The worth of the feeling might be favourable or negative. Understanding the statement's intent is important in order to identify the phrase's implications and the node of action to which the statement refers for this, knowledge in text processing and analysis is needed.

While a person may arrive at the conclusion quickly, a machine would rather take a long time to get there. One person struggles to assess the sentiment of the communications when there are more tweets and papers to analyze. We anticipate that machines will be able to process data fast as a consequence.

1.1.1 Importance of sentiment analysis in business

In business sector have some benefits using sentiment analysis like customers gives their feedback based on their requirements and it is very important thing to increase their product quality or quantity as well as business in market. It is helped to develop their business in all departments. By the sentiment analysis the company can detect the inner feeling about the company, its services, and its products from the customers.

In the business sector, there are some

key processes to sentiment analysis: obtaining data, such as customer feedback, cleaning text by removing stopwords, evaluating the data, and comprehending the findings. the business must comprehend client sentiment, and it will be able to accomplish so by constantly observing customer responses.

2. Methodology

First we have some different methodology in sa they are document-level sentiment analysis, topic-based sa and aspect-based sentiment analysis this research project comes under document-level sa In this the people feelings can represented as positive as 1, negative as 0.

1.1 methods to analyze the sentiment data

Data Collection

In public forums like blogs, message boards, and product reviews as well as on personal accounts like Facebook and Twitter, consumers frequently express their thoughts. Opinions and opinions are expressed in a variety of ways, using different vocabulary, writing contexts, and slang and short forms, which creates a huge and disorganised data collection. It is nearly impossible to manually analyse sentiment. As a result, certain programming languages, such "R," are used to handle and analyse the data.

Text preparation

Text preparation is the process of filtering the collected data before analyzing it. It entails finding and deleting from the data any non-textual and non-research-related information.

Sentiment detection

At this step, the objectivity of each review and opinion sentence is examined. Sentences with subjective phrases are retained, while those with objective ones are omitted. By using well-known computing techniques like lemmas, negation, and unigrams, sentiment analysis is carried out at several levels.

Sentiment classification

Positive(1) and negative(0) emotions can be classed widely. Each detected subjective remark is categorized as positive, negative, good, bad, like, or dislike at this step of the sa process.

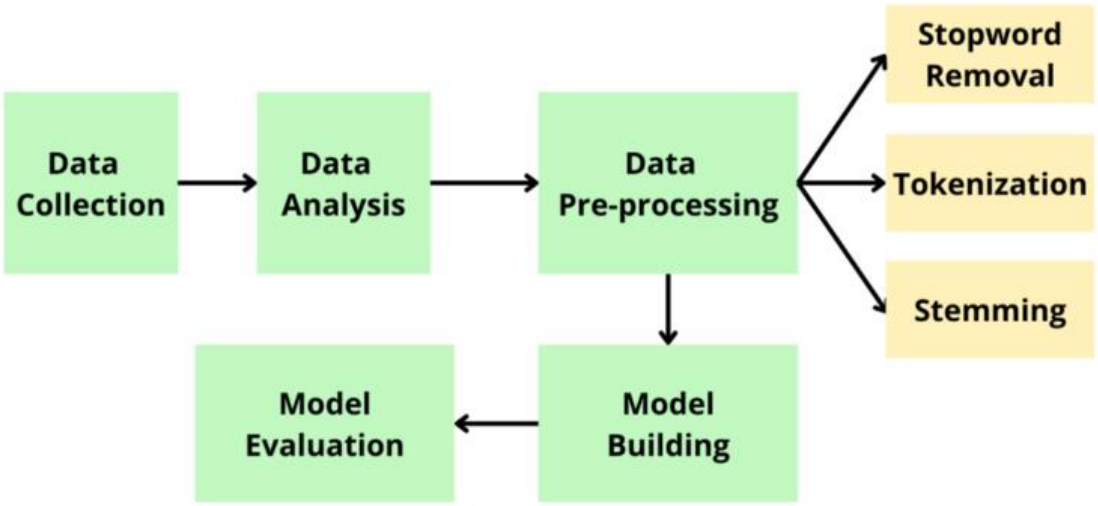
Result or output

Making shapeless data into useful knowledge is the main important thing of sa. The text findings are shown on graphs, such as pie charts, when the analysis is complete.

Using Naive Bayes

For huge amounts of data, Nb is the simplest and quickest categorization approach. The Nb classifier is widely used in applications such as spam filtering, text categorization, sa, and recommendation systems. The Bayes probability theorem is used to forecast unknown classes. In ml, the Nb technique is a simple and efficient classification problem.

The Bayes theorem and a strong conviction in feature independence are the basis of naive Bayes classification. The Nb classification approach works well with textual data analysis, such as that involved in NLP. Other names for them include simple Bayes models, independent Bayes models, and naïve Bayes models. All of them are related to the classifier's Bayes theorem-based decision-making process.



Logistic Regression

Any word or piece of data may be expressed as a vector of dimension V, where V is the same as the scope of our vocabulary. You would put a 1 in the associated index for each phrase in the tweet if you received the data "I am researching sentiment analysis," for example, and a 0 if you didn't. As V increases, the vector, as can be seen, becomes sparser. We also have to train V parameters since we end up with a lot more features. Longer training and prediction times may be the outcome of this.

As a result, we will extract frequencies from each word and create a frequency dictionary. The goal is to separate the training set into positive and negative feedback. Count the number of words.

Logistic regression is a technique used to forecast the outcome of a categorical dependent variable. As a result, the conclusion must be distinct or categorical. It provides probability values between 0 and 1 rather than precise integers between 0 and 1. It might be True, False, Yes, No, 0 or 1, etc.

Discussion

Figure 1. [Review Result]

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

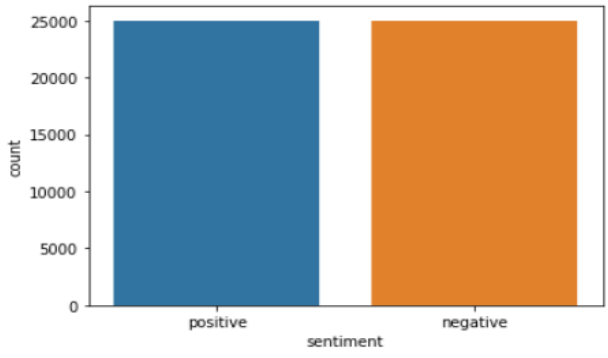
Here in figure 1 represents the result of the given movie review like positive, negative.

Figure 2. [Result in digits]

	review	sentiment
0	One of the other reviewers has mentioned that ...	1
1	A wonderful little production. The...	1
2	I thought this was a wonderful way to spend ti...	1
3	Basically there's a family where a little boy ...	0
4	Petter Mattei's "Love in the Time of Money" is...	1

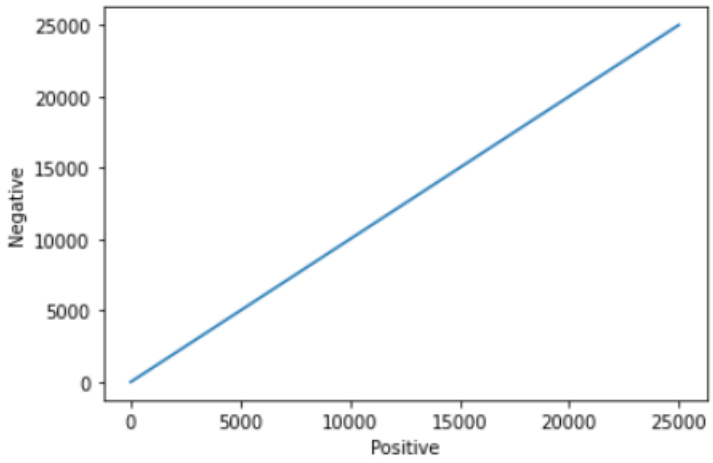
Here in figure 2 represents the result of the given movie review in numerical like positive-1, negative-0.

Figure 3. [Total reviews]



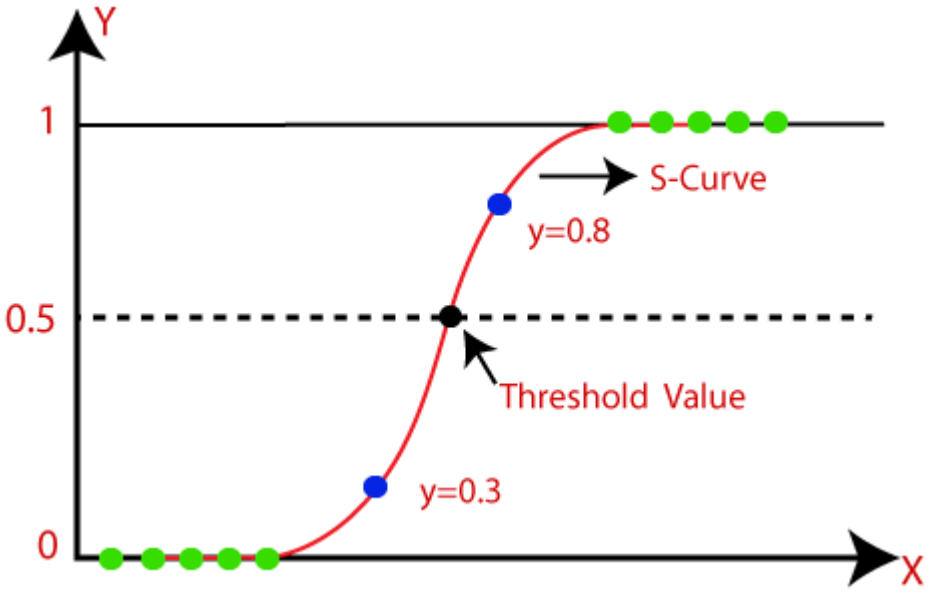
Here in figure 3 represents the total reviews taken like
Total=50,000;Positive=25,000;Negative=25,000

Figure 4. [Total reviews in graph]



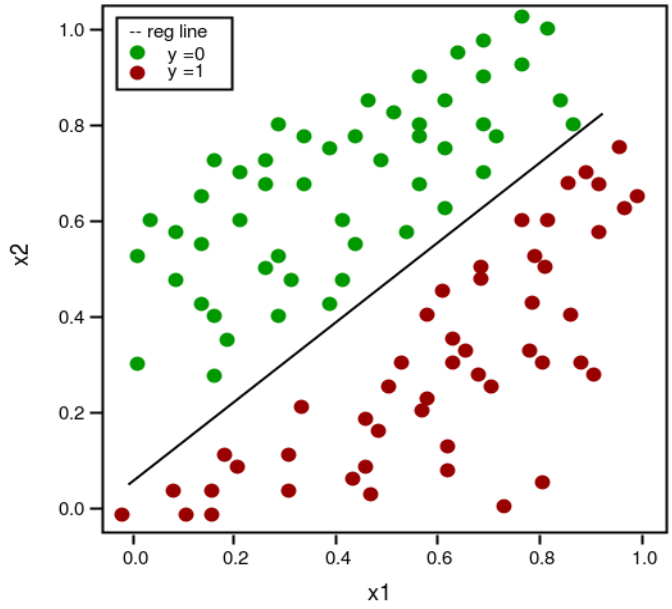
Here in figure 4 shows the total reviews in graph

Figure 5. [Logistic Regression]



Here in figure 5 represents the graph of logistic regression

Figure 6. [Graph]



Here in figure 6 represents the total(positive, negative) in divided form.

Table 1. [Result Table]

Reviews	Result	Result
This is the wonderful movie of my life	Positive	1
It is very bad movie	Negative	0
It was a best movie	Positive	1

Table 2. [Actual, predicted Table]

10000 rows * 2 columns

	Actual	predicted
0	1	1
1	1	1
2	1	1
3	1	1
…….	…….	…….
9998	0	0
9999	0	0

Table 3. [Precision Table]

	Precision	Recall	F1-score	support
0	0.85	0.85	0.85	4959
1	0.85	0.85	0.85	5041
Accuracy			0.85	10000
Macro avg	0.85	0.85	0.85	10000
weighted avg	0.85	0.85	0.85	10000

Table 4. [Accuracy Table]

Algorithms	Data Size	Accuracy
Naive Bayes	50000	0.85=85%
Logistic Regression	100000	0.77=75%

Equation 1. [Naive Bayes]

$$P(A \mid B) = P(B \mid A) * P(A) / P(B)$$

P(A|B) is posterior likelihood

P(B|A) is Probability of occurrence

P(A) is Probability of Priority

P(B) is Probability at the margins

Equation 2. [Logistic Regression]

$$Y=e^{(b0+b1x)}/1+e^{(b0+b1x)}$$

x = value that given to input

y = the expected output value.

b0 = word for bias or intercept.

b1 = a factor for input (x)

Concluding Remarks

The project's text representation strategy was the bag of words technique. Finally, the results were obtained using two different traditional machine learning techniques. The two models we used are naive Bayes and logistic regression. Our models are trained and built on the Kaggales IMDB dataset. The Naive Bayes classifier with its feature set gives us the best accuracy. Aside from that, we can employ the Logistics Regression Classifier. We discover that the Navy's Bayes model performs the best, with an accuracy of 0.87. The accuracy for logistic regression is then 0.77.

Future Work

Till now, based on the given time we have used some basic procedures and algorithms with 85 percent accuracy which can be improved and enhanced by using different algorithms and combinations. This task done by the team is just to analyze the sentiment between positive and negative only, it can be further improved. So, it can also be used to find neutral sentiments and sentiments based on emoji's

References

1. data set from Kaggle
2. Navie bayes
3. natural language process
4. stopword
5. logistic regression