Sentimental Analysis using IMDB Reviews

Akanksha Porwal 202118017
Dipshi Jain 202118018
Jenil Doshi 202118034
Rachit Kumar Singh 202118015

April 2022

1 Abstract

In this project, we have worked on the sentiment analysis of IMDB Reviews with NLP. In particular, we have predicted the number of positive and negative reviews based on sentiments by using different classification models such as logistic regression,multinomial naive bayes model and linear support vector machines.In this project, we are comparing accuracy of these models and observing the variations in results given by these three models which we will discuss in this report.

2 Introduction

Sentiment analysis is used to analyse a given statement's actual meaning or effect.It's also known as opinion mining.It combines machine learning and natural language processing (NLP) to achieve this. Basically,there is a sentiment analysis model that classifies the text into positive or negative sentiments.Here we make use of supervised learning models where we provide fair amount of labelled data to the model for training.These supervised learning models basically gives output in numbers which indicates how similar the text is to the positive text or to the negative text during the training.

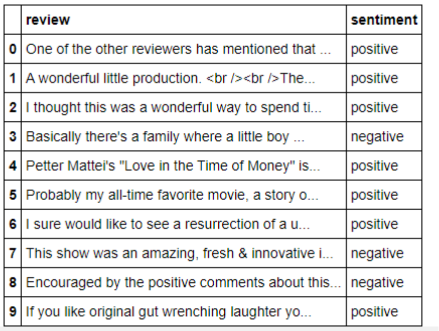
For this project we have used the well-known IMDB dataset.It was released to the public by Stanford University.This dataset is a collection of 50,000 reviews from IMDB that contains an even number of positive and negative reviews.The dataset have no more than 30 reviews per movie.Considering the scores for positive and negative review except for neutral reviews, a bad review has less than or equal to 4 out of 10 and for positive review it is greater than or equal 7 out of 10.

The dataset was compiled by Andrew Maas and initially introduced in this paper: Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis.

3 Dataset

The below figure is snapshot of our dataset showing first 10 enteries.For the sentiment analysis the dataset the

IMDB dataset that we are using has 50k movie reviews.It consist of 25,000 movie reviews for training and 25,000 movie reviews for testing. So,we will predict the number of positive and negative reviews using some classification algorithms.



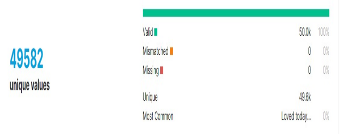
	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. <br /><br />The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive
5	Probably my all-time favorite movie, a story o...	positive
6	I sure would like to see a resurrection of a u...	positive
7	This show was an amazing, fresh & innovative i...	negative
8	Encouraged by the positive comments about this...	negative
9	If you like original gut wrenching laughter yo...	positive

Raw Data with first 10 entries

If we talk about any organization we can observe that sentiment analysis helps to understand user sentiments and opinions for a particular service and product. It also heavily impacts improving the business logic and overall profit of an organization by bringing what their customers prefer.

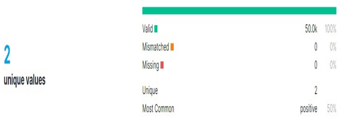
Here,we tried analysis by first doing text preprocessing using the standard NLP methods and applying Language Understanding Algorithms or model to predict user sentiments.

Also, in this dataset 50,000 movie reviews are pre-labeled with “postive” and “negative” sentiment class labels.



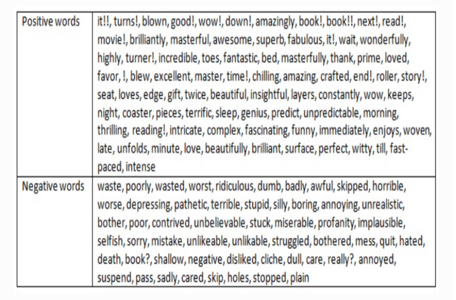
49582
unique values

Valid 🟩	50.0k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	49.6k	
Most Common	Loved today...	0%



2
unique values

Valid 🟩	50.0k	100%
Mismatched 🟧	0	0%
Missing 🟥	0	0%
Unique	2	
Most Common	positive	50%



Positive words	it!!, turns!, blown, good!, wow!, down!, amazingly, book!, book!!, next!, read!, movie!, brilliantly, masterful, awesome, superb, fabulous, it!, wait, wonderfully, highly, turner!, incredible, toes, fantastic, bed, masterfully, thank, prime, loved, favor, !, blew, excellent, master, time!, chilling, amazing, crafted, end!, roller, story!, seat, loves, edge, gift, twice, beautiful, insightful, layers, constantly, wow, keeps, night, coaster, pieces, terrific, sleep, genius, predict, unpredictable, morning, thrilling, reading!, intricate, complex, fascinating, funny, immediately, enjoys, woven, late, unfolds, minute, love, beautifully, brilliant, surface, perfect, witty, till, fast-paced, intense
Negative words	waste, poorly, wasted, worst, ridiculous, dumb, badly, awful, skipped, horrible, worse, depressing, pathetic, terrible, stupid, silly, boring, annoying, unrealistic, bother, poor, contrived, unbelievable, stuck, miserable, profanity, implausible, selfish, sorry, mistake, unlikeable, unlikable, struggled, bothered, mess, quit, hated, death, book?, shallow, negative, disliked, cliche, dull, care, really?, annoyed, suspend, pass, sadly, cared, skip, holes, stopped, plain

Collection of Sentimental Words

4 Process

In this context, we will be focusing on preparing the data set, eliminating the redundancies such as punctuation’s, and stop words etc and then using different classification models to predict the sentiments.

The steps to be followed can be proposed:

1. Loading the data set
2. Preprocessing of data
   (a) Lemmatizing
   (b) Tokenizing
3. Cleaning
   (a) Stopword
   (b) Punctuation’s
   (c) Common words
   (d) Brackets
   (e) HTML tags
   (f) Emojis

The above steps given as a basic necessity for any pipeline which is related to NLP(Natural language processing),the transformed data is then used for analysis. After, this we call our data normalized and fit for passing to any classification based model.Also we can try to fine tune the data with different-different transformations.

5 Data-Preprocessing

lines cover the outline of what we did to normalize our dataset-

1. Text Normalization
   Words are tokenized. To separate a statement into words, we utilise the word tokenize () method.

2. Removing html strips and noise text

   Here in data if we have some html code, we need to clean that html strips. Also removing some noisy texts along with square brackets.

3. Removing special characters

   As our dataset is in English-language we need to make sure that any special characters are deleted.

4. Text stemming

   Stemming is a technique for eliminating affixes from words in order to retrieve the base form or word-stem. The stem need not be identical to the morphological root of the word.It's the same as pruning a tree's branches down to the trunk. The stem of the terms eating, eats, and eaten, for example, is eat.

5. Removing stop words and normalization

   Stop words are words that have little or no meaning,these are most common words in any natural language.For the purpose of analyzing movie reviews, these stop words might not add much value to the meaning of the document.

   Generally, the most common ones are "the", "is", "in", "for", "where", "when", "to", "at" etc.It is very important to filter out of natural language data before or after it is processed in computers. We consider this text string – "There is a pen on the table". Now, the words "is", "a", "on", and "the" add no meaning to the statement while parsing it. Whereas words like "there", "book", and "table" are the keywords and tell us what the statement is all about.

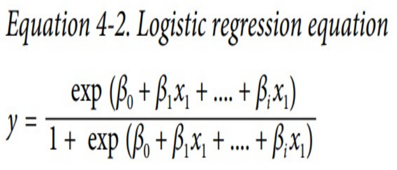
   Text normalisation in sentiment analysis is the process of cleaning or removing irrelevant data from a huge collection of extracted data. Because input is guaranteed to be consistent before operations are done on it, normalising text before storing or processing it allows for separation of concerns.

6 Models used

6.1 Logistic Regression

After analysing and monitoring the data, logistic regression is one of the most extensively used algorithms for classification. It predicts a binary result such as yes or no, True or False, and so on. It assures that the output probabilities add up to one and stay between zero and one, as we would anticipate. The logistic regression model (or logit model) is a variant of linear regression that employs the sigmoid function.

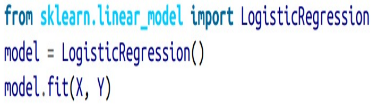
A logistic regression model analyses the connection between one or more existing independent variables to predict a dependent data variable. A logistic regression, for example, might be used to predict whether a political candidate will win or lose an election, or if a high school student would be admitted or not to a specific institution.



Equation 4-2. Logistic regression equation

$$y = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_i x_1)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_i x_1)}$$

Logistic Regression Equation



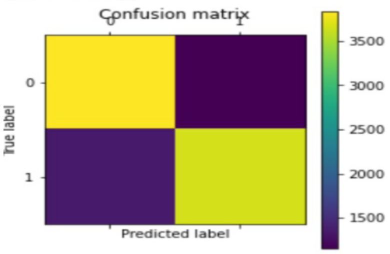
```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X, Y)
```

Python Code

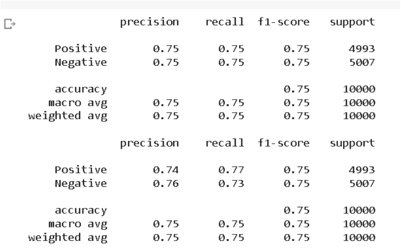
In the diagram below, y represents the expected output, o represents the bias or intercept term, and B1 represents the coefficient for a single input value (x). Each column in the input data has a corresponding coefficient (a constant real value) that must be determined using the training data.

Maximum likelihood estimation (MLE) approaches are used to forecast values near to 1 for the default class and close to 0 for the other class while training the logistic regression coefficients.

We have used logistic regression model using Logistic Regression class of the sklearn package of Python.



Confusion Matrix

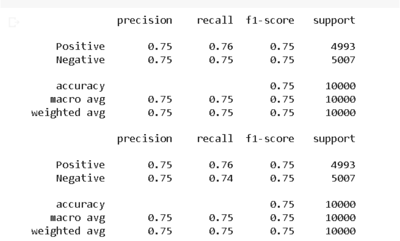


```
              precision    recall  f1-score   support

    Positive       0.75      0.75      0.75      4993
    Negative       0.75      0.75      0.75      5007

    accuracy                           0.75     10000
   macro avg       0.75      0.75      0.75     10000
weighted avg       0.75      0.75      0.75     10000
              precision    recall  f1-score   support

    Positive       0.74      0.77      0.75      4993
    Negative       0.76      0.73      0.75      5007

    accuracy                           0.75     10000
   macro avg       0.75      0.75      0.75     10000
weighted avg       0.75      0.75      0.75     10000
```

Classification Report

6.2 MultinomialNaive Bias Classifier

The Naive Bayes technique is an effective way for assessing text input and solving issues involving several classes. Because the Naive Bayes theorem is based on the Bayes theorem, a basic understanding of the Bayes theorem is required. The Bayes theorem, devised by Thomas Bayes, calculates the probability of an event occurring based on prior knowledge of the event's circumstances. We calculate the chance of class A when predictor B is provided. It's calculated using the following formula: P(A—B) = P(A) * P(B—A)/P (B).

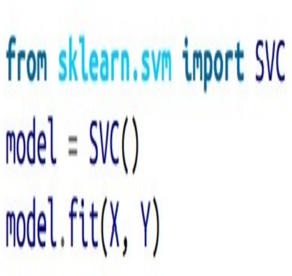


```
              precision    recall  f1-score   support

    Positive       0.75      0.76      0.75      4993
    Negative       0.75      0.75      0.75      5007

    accuracy                           0.75     10000
   macro avg       0.75      0.75      0.75     10000
weighted avg       0.75      0.75      0.75     10000
              precision    recall  f1-score   support

    Positive       0.75      0.76      0.75      4993
    Negative       0.75      0.74      0.75      5007

    accuracy                           0.75     10000
   macro avg       0.75      0.75      0.75     10000
weighted avg       0.75      0.75      0.75     10000
```

Classification Report

6.3 SVM Classifier

Support vector machines (SVMs) are a family of supervised learning algorithms for classification, regression, and identification of outliers. The goal of the support vector machine (SVM) method is to maximise the margin, which is defined as the distance between the separating hyperplane (or decision border) and the nearest training samples, also known as support vectors. We use the perpendicular distance from the line to just the nearest spots to determine the margin.



```
from sklearn.svm import SVC
model = SVC()
model.fit(X, Y)
```

Python Code

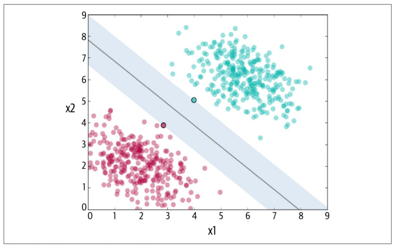
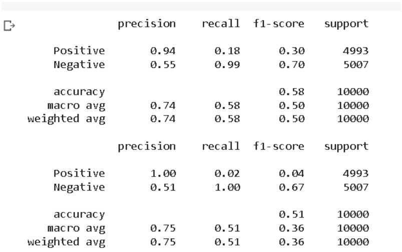


Figure 4-3. Support vector machine

Support Vector Machine



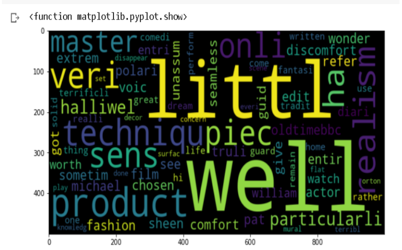
```
              precision    recall  f1-score   support

    Positive       0.94      0.18      0.30      4993
    Negative       0.55      0.99      0.70      5007

    accuracy                           0.58     10000
   macro avg       0.74      0.58      0.50     10000
weighted avg       0.74      0.58      0.50     10000
              precision    recall  f1-score   support

    Positive       1.00      0.02      0.04      4993
    Negative       0.51      1.00      0.67      5007

    accuracy                           0.51     10000
   macro avg       0.75      0.51      0.36     10000
weighted avg       0.75      0.51      0.36     10000
```

Classification Report

Some of the advantages of support vector machines are it is effective in high dimensional spaces.It is versatile as different kernel functions can be specified for the decision function.Also,it uses a subset of training points in the decision function (called support vectors), so it is also memory efficient

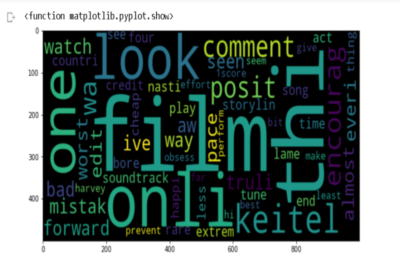
7 NLP

A formal definition of NLP typically includes terminology to the effect that it is an area of research that analyses natural language utilising computer science, artificial intelligence, and formal linguistics ideas. It is a set of methods used to derive meaningful and usable information from natural language sources such as web pages and text documents, according to a less formal definition. NLP methods are used to process a user query and provide a result page that the user may utilise. When working with a language, we regularly come across terminology, syntax, and semantics. The rules that regulate a legitimate sentence construction are referred to as a language's syntax. For example, in English, a common sentence structure begins with a subject, then a verb, and finally an object, as in ”Tim hit the ball.” We aren't used to sentences in unexpected sequence, such as ”Hit ball Tim.” Despite the fact that English syntax is less strict than that of computer languages, we nonetheless require a phrase to follow fundamental syntactic principles. The meaning of a statement is defined by its semantics. We comprehend the meaning of the phrase ”Tim hit the ball” as English speakers. However, English and other natural languages are sometimes ambiguous, and the meaning of a statement may only be deduced from its context.



```
<function matplotlib.pyplot.show>
```

Word Cloud of Positive Words



```
<function matplotlib.pyplot.show>
```

Word Cloud of Negative Words

8 Conclusion

We can observe that both logistic regression and multinomial naive bayes model performing well compared to linear support vector machines.

9 Future Work

Opinion Mining or Sentiment Analysis on a group of reviews and identified feature expressions taken from reviews will be possible in the future. Rather than being good or negative, we may try to anticipate other attitudes behind the reviews, such as toxic, severe toxic, obscene, threat, insult, and identity hate.

We can also try to apply different vectorizations in the future to improve the word matrix. For example, we may try removing the subjects from the phrases. In addition, future research might test more complex models for analysis. For example, because it can account for the link between the sentences, a recurrent neural network may be able to deliver superior results.

10 References

1. K. Amulya, S. B. Swathi, P. Kamakshi and Y. Bhavani, ”Sentiment Analysis on IMDB Movie Reviews using Machine Learning and Deep Learning Algorithms,” 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), 2022, pp. 814-819, doi: 10.1109/ICSSIT53264.2022.9716550.

2. S. Tripathi, R. Mehrotra, V. Bansal and S. Upad-

hyay, "Analyzing Sentiment using IMDb Dataset," 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), 2020, pp. 30-33, doi: 10.1109/CICN49253.2020.9242570.

3. T. P. Sahu and S. Ahuja, "Sentiment analysis of movie reviews: A study on feature selection classification algorithms," 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), 2016, pp. 1-6, doi: 10.1109/MicroCom.2016.7522583.

4. N. L. Adam, N. H. Rosli and S. C. Soh, "Sentiment Analysis on Movie Review using Naïve Bayes," 2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS), 2021, pp. 1-6, doi: 10.1109/AiDAS53897.2021.9574419.