Sentiment Analysis

M. NARASIMHA RAO - AP20110010017

B. JAYANTH - AP20110010049

G. HARSHA VARDHAN - AP20110010056

E. VARA SIDDHA VIGNESH – AP20110010058

Outline

- ❑ Introduction
- ❑ Need of Sentiment Analysis
- ❑Approach for Sentiment Analysis
- ❑ Implementation
- ❑ Advantages
- ❑ Conclusion

What is Sentiment Analysis

- Sentiments are feelings, opinions, emotions, likes/dislikes ,good/bad.
- Sentiment Analysis is a Natural Language Processing and Information Extraction task that aims to obtain writer's feeling expressed in positive or negative comments, questions and request, by analyzing a large numbers of documents.
- Sentiment Analysis is a study of human behavior in which we extract user opinion and emotion from plain text.
- Sentiment Analysis is also known as Opinion Mining.

What is Sentiment Analysis

- It is a task of identifying whether the opinion expressed in a text is positive or negative.
- Automatically extracting opinions, emotions and sentiments in text
- Language-independent technology that understand the meaning of the text.
- It identifies the opinion or attitude that a person has towards a topic.

Example

- ❑ **User's Opinions:**

  Jayanth: It's a great movie (Positive statement)

  Vardhan: Nah!! I didn't like it at all (Negative statement)

  Vignesh: The new Avatar movie is awesome..!!!(Positive statement)

- ❑ **Polarity:**

- Positive
- Negative

Approach

❑ **NLP(Natural Language Processing)**

- Use semantics to understand the language.

❑ **Machine Learning**

- Don't have to understand the meaning
- Uses classifiers such as Naïve Byes, Logistic Regeression.

Naïve Bayes

❑The Naive Bayes algorithm is a supervised machine learning algorithm based on the Bayes’ theorem. It is a probabilistic classifier that is often used in NLP tasks like sentiment analysis.

❑Advantages of working with NB algorithm are:

- Requires a small amount of training data to learn the parameters Can be trained relatively fast compared to sophisticated models

❑The main disadvantage of NB Algorithm is:

- It’s a decent classifier but a bad estimator It works well with discrete values but won’t work with continuous values

Logistic regression

❑ Logistic regression is one of the most popular Machine Learning algorithms.

- It is used for predicting the categorical dependent variable using a given set of independent variables . Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either 0 or 1, true or False, etc.

❑ Steps in Logistic Regression:

- Data Pre-processing step
- Fitting Logistic Regression to the Training set
- Predicting the test result
- Test accuracy of the result
- Visualizing the test set result.

Equations

Naïve Bayes

$P(A|B) = P(B|A) * P(A) / P(B)$

$P(A|B)$ is posterior likelihood

$P(B|A)$ is Probability of occurrence

$P(A)$ is Probability of Priority

$P(B)$ is Probability at the margins

Logistic Regression

Y=e^(b0+b1x)/1+e^(b0+b1x)

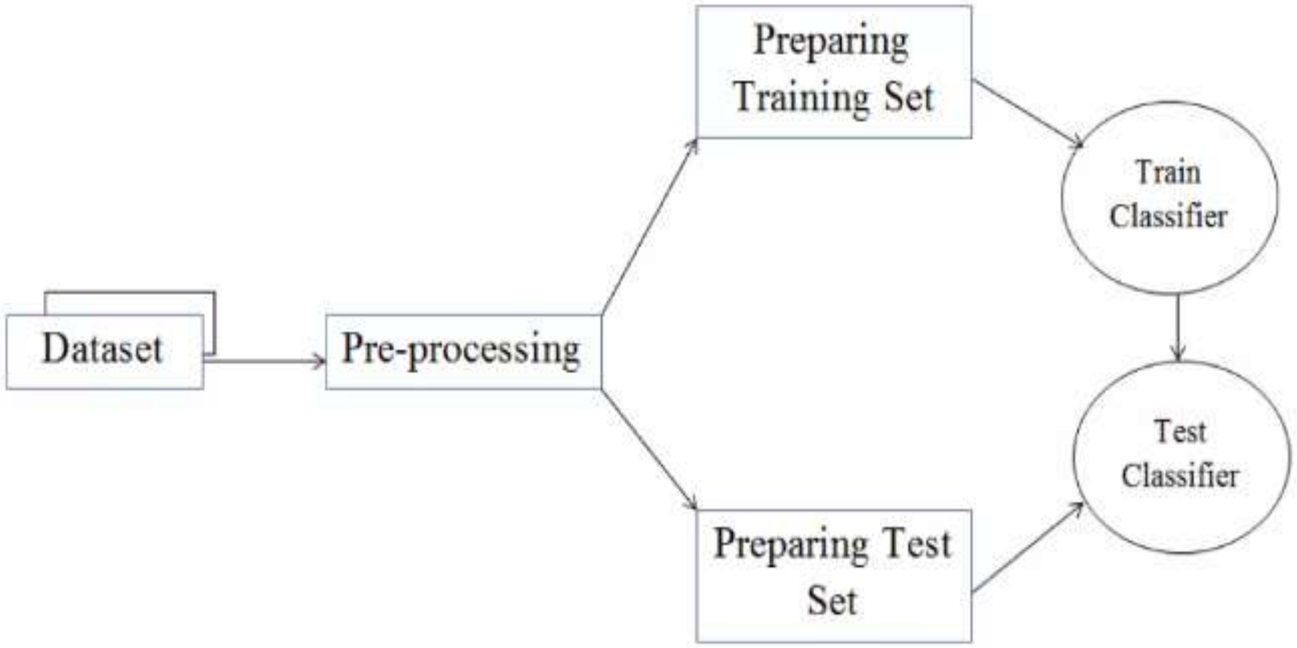
x = value that given to input

y = the expected output value.

b0 = word for bias or intercept.

b1 = a factor for input (x)

Implementation



Pre-Processing

❖**Tokenization**

❑Unigram: considers only one token

- e.g. It is a good movie.
- (It, is, a, good, movie}

❑Bigram: considers two consecutive tokens

- e.g. It is not bad movie.
- {It is, is not, not bad, bad movie}

Pre-Processing

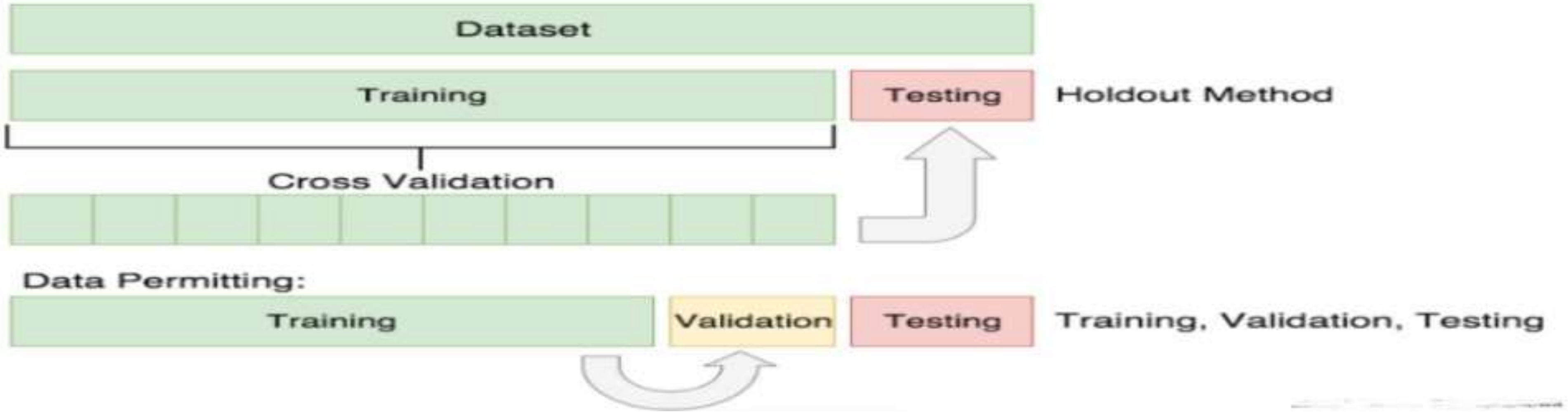
- ❑ **Stopwords Removal** - Common words such as "is", "am", "are", "the" etc. are likely to give no meaning to the text. These words are used just to help the main meaning giving words. Such words are called stopwords.

- ❑**Text Normalization-** Movie reviews are generally a form of casual writing. For example, "good" at times may be written as "goooood", "gud", "gd" etc. All of these words will impart similar meaning but are available in different forms.

Pre-Processing

Train Test Split

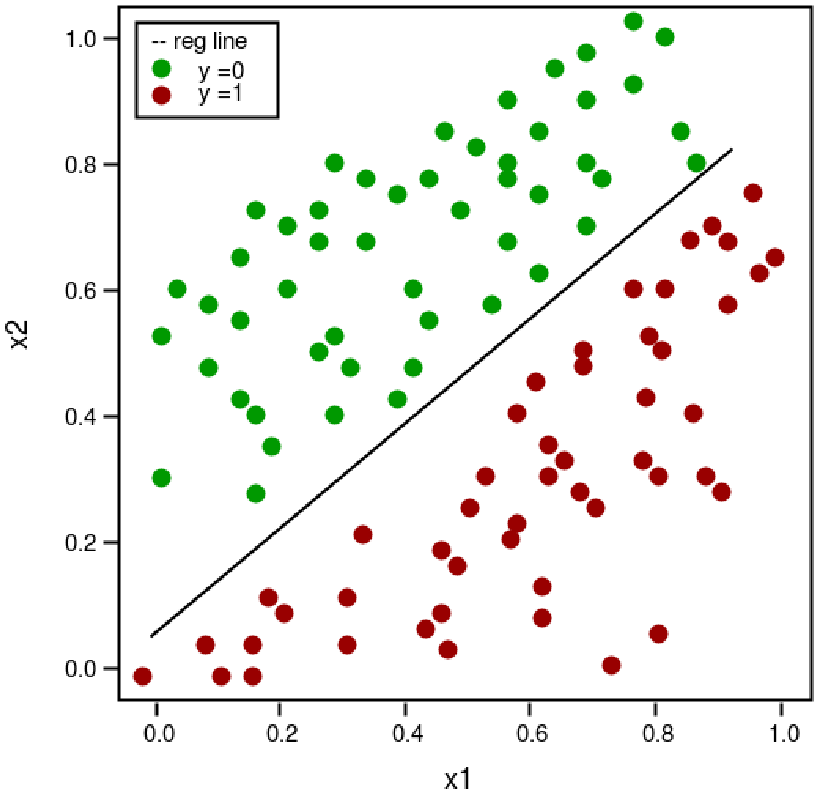
- To measure the accuracy of the model we are creating, the data needs to split into 2 parts. A training set to fit and tune our model and a testing set to create predictions on and evaluate the model at the very end.



Accuracy

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

Represents the total(positive, negative) reviews



Advantages

- A lower cost than traditional methods of getting movie reviews.
- A faster way of getting insight from movie lovers.
- The ability to act on viewers suggestions.
- Identifies an organization's Strengths, Weaknesses, Opportunities & Threats .
- As 80% of all data for a movie review consists of words, the Sentiment Engine is an essential tool for making sense of it all.
- More accurate and insightful movie reviews and feedback.

Conclusion

- The project's text representation strategy was the bag of words technique.
- The two models we used are naive Bayes and logistic regression.
- 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

Bibliography

- https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
- https://web.stanford.edu/~jurafsky/slp3/4.pdf
- https://stanfordnlp.github.io/CoreNLP/index.html
- https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a
- https://towardsai.net/p/nlp/sentiment-analysis-with-logistic-regression

THANK YOU