Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

Comparative Analysis of Book Reviews using SVM Linear and RBF Kernel

Om Kolte
Department of Computer Engineering,
Pimpri Chinchwad College of Engineering
Pune, India

Prof. Archana Kadam
Department of Computer Engineering,
Pimpri Chinchwad College of Engineering
Pune, India

***Abstract*— Sentiment analysis, also referred to as opinion mining, is a branch of natural language processing which focuses on the analysis of identifying the opinions or feelings expressed in textual content. The primary focus of this study is on conducting a sentiment classification of book reviews using supervised (Support Vector Machine) machine learning technique on a book review dataset from Amazon. The comparative analysis of the approach with different kernel parameter indicates that supervised approach (SVM) with Radial Basis Function (rbf) kernel gives the best accuracy of 83.84% whereas linear kernel gives us better overall result with accuracy of 81.24%.**

Keywords—Sentiment analysis, Support Vector Machine, Supervised approach, Machine Learning, Radial Basis Function (rbf)

I. INTRODUCTION

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data means some input data is already tagged with the correct output. This includes algorithms like Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes Classifier (NB), and many more.

Unsupervised learning is the type of machine learning in which models are not supervised using training dataset. Its main goal is to find hidden patterns from the data on which it is applied. This includes algorithms like the semantic oriented approach (SO-PMI-IR).

The introduction of the internet and technology has provided people with greater access to web apps via smart devices, greatly increasing the importance of the rating system. However, there are millions of product or service related reviews available on the web, and reading all of them is a time-consuming and stressful effort for anyone. As a result, there is a need for appropriate approaches that automatically categorize these reviews as good or negative in order to provide valuable information to the user. This classification task is technically known as Sentimental Analysis. It is a branch of natural language processing which focuses on the analysis of identifying the opinions or feelings expressed in textual content. Data mining is a process that helps extract useful knowledge from large amounts of data. Analysis of sentiment is performed constantly in widely spoken languages such as English with various methods of machine learning algorithms'.

II. LITERATURE REVIEW

The supervised and unsupervised approaches used for opinion mining were discussed. The supervised approach consists of Naïve Bayes Approach, Support Vector Machine (SVM) and the unsupervised approach consists of the semantic approach also known as SO-PMI-IR algorithm used for classifying reviews. The results of this paper shows that the unsupervised algorithms works better with dataset having long phrases whereas the supervised algorithms gave higher accuracy on the dataset containing one-lined short reviews. V. Kaur proposed an unsupervised semantic oriented approach for classifying books as positive, negative or neutral based on the reader's reviews of the respective books. The algorithm which was used is called as SO-PMI-IR where:

- SO – Semantic Orientation
- PMI – Pointwise Mutual Information
- IR – Information Retrieval

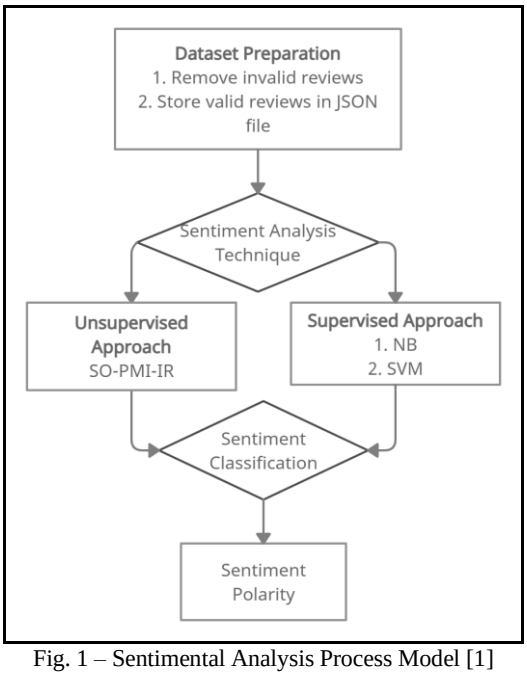

Fig. 1 – Sentimental Analysis Process Model [1]

The first phase, which is depicted in figure 1, is the preparation of the dataset, in which empty and incomprehensible reviews are deleted from the dataset, and then the text of processed reviews is retrieved from those reviews and stored in text file. Then, in the second stage, for the purpose of sentiment classification, two supervised techniques known as NB and SVM as well as an unsupervised approach known as SO-PMI-IR were utilised.

The comparison of Naïve Bayes Algorithm with the results of the semantic approach SO-PMI-IR and SVM was analyzed.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

It is observed that better accuracy is achieved with NB rather than SVM and is almost similar to the SO-PMI-IR approach. The results showed that SO-PMI-IR gave the best accuracy and NB outperformed SVM.
P. Walia explored unsupervised SO-PMI-IR as well as the supervised NB and SVM approaches for sentimental analysis of the reviews. The results showed that SO-PMI-IR gave the best accuracy and the NB classifier outperformed the SVM. [2]

The provision of a visual method for book review evaluations in order to assist readers in gaining knowledge of various books. The purpose of the suggested system is to offer a graphical method for reviewing books, with the intention of assisting users in gaining a better knowledge of a variety of works through the utilization of various visuals. Through the use of Tableau software, this paper reported the findings of a study that analyzed 1000 customer reviews posted on Amazon. [3]

Multiple methodologies used for pre-processing techniques. Analysis of user sentiment and a representation of various visualization approaches make up the two primary components that are integrated into the system to provide answers to the research questions. [4]

Using a clustering algorithm for review classification which is used to group the users into clusters of their interests and collaborative algorithm is used to recommend books. There are multiple algorithms implemented for opinion mining. [5]

III. ALGORITHMIC SURVEY

1. Support Vector Machine:

The Support Vector Machine, sometimes known as SVM, is one of the most common and widely used supervised learning algorithms. Its primary function in machine learning is to solve classification problems. The purpose of SVM is to generate the best line or decision boundary that can divide an n-dimensional space into classes. This will allow us to easily place any new data points in the appropriate category in the future. A hyperplane is the term used to describe this optimal decision boundary.
The extreme points and vectors that contribute to the creation of the hyperplane are selected using SVM. These exceptional circumstances are referred to as support vectors, which is how the method got its name: the Support Vector Machine. Take a look at the diagram below, which shows how two distinct groups can be differentiated from one another with the help of a decision boundary or a hyperplane:

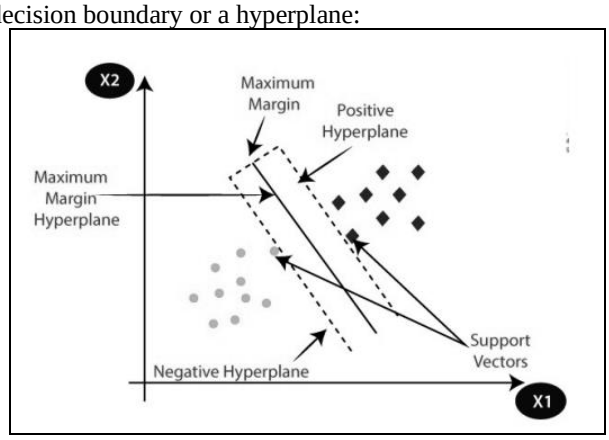

Fig. 2 – Support Vector Machine (SVM)

There are two different types of SVMs, and both are used for different things:

i. Simple SVM
   a. Used for linear regression and classification problems
ii. Kernel SVM
   a. Provides more flexibility for non-linear data
   b. Allows to add more features to fit a hyperplane instead of 2D space.

2. Pointwise Mutual Information:

In the fields of information theory and statistics, a measure of association called pointwise mutual information (PMI), commonly known as point mutual information, is used. Pointwise Mutual Information (PMI) between two adjectives is as follows:

$$PMI(x,y) = log_2(\frac{P(x,y)}{P(x)P(y)}). \quad (1)$$

where the probability that word1 and word2 occur together is denoted by p(x,y), the probability that word1 appears by itself is denoted by p(x), and the probability that word2 appears by itself is denoted by p(y). If there is a genuine connection between word1 and word2, then the probability that occur in p(word1& word2) will be significantly higher than the odds that occur in p(word1)p(word2). The logical conclusion is that the value of log(word1 word2) is greater than 0. The amount of information that can be gleaned about the presence of one word by observing the presence of another word is denoted by the log of this ratio. One way to determine the orientation of an adjective is to use the SO-PMI-IR notation.
SO-PMI-IR(word) = PMI(word, {positive_paradigms}) – PMI(word, {negative_paradigms})
If PMI result is positive then adjective word has positive semantic orientation.
If PMI result is negative then adjective word has negative semantic orientation.

3. Performance Measures:

The following measures have been used to analyze the performance of the algorithm used:

i. **Accuracy = (True Positive + True Negative) / Total instances**
ii. **Precision = True Positive / (True Positive + False Positive)**
iii. **Recall = True Positive / (True Positive + False Negative)**
iv. **F-score = 2. (Precision. Recall/ (Precision + Recall))**

The techniques used in this paper have been implemented in Python and the dataset was in JSON format.

IV. PROPOSED SYSTEM

Going through the existing research papers in this domain of research and their methodology, we have tried to propose a system which mainly involves the following 5 stages:

Stage 1 – Dataset Preparation
Stage 2 – Dataset Preprocessing
Stage 3 – Feature Extraction

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

Stage 4 – Classification
Stage 5 – Performance Evaluation and Comparison

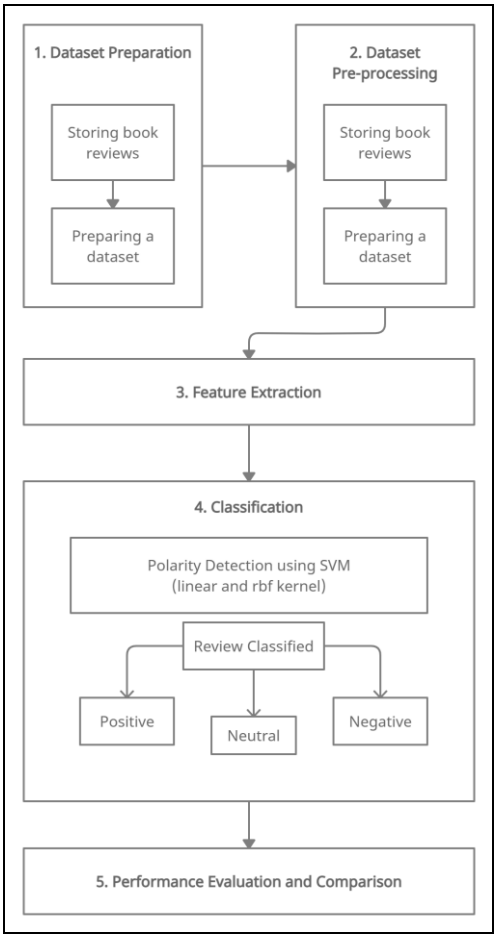

Fig. 3 – Proposed System Model

Stage 1 – Dataset Preparation

This stage involves collecting book reviews from Amazon Marketplace and storing them into a file to create a dataset. This file can be of any type like csv, json, etc. The main aim of this stage is to collect the data and store it in a file for further processing and use.

Stage 2 – Dataset Preprocessing

In this stage, the data which was collected and stored is then preprocessed. The dataset contains reviews with null values or some invalid values, these are removed from the dataset and then the data is stored in the file again after all those values are removed. This gives us a clean dataset which is easy to use and would yield better results. Basically, the null and invalid entries are removed from the original dataset to obtain a clean dataset.

Stage 3 – Feature Extraction

The approach known as dimensionality reduction, in which an initial collection of unprocessed data is partitioned into categories that are easier to work with, includes the feature extraction step as one of its components. Because of this, the processing will be made easier. Feature extraction is something that we do since it makes it possible for the model to be produced with less work from the machines and speeds up the processes of machine learning and generalization.

Stage 4 – Classification

After the data has been pre-processed, we apply the SVM algorithm to identify the polarity of the review. First we used the rbf kernel optimization parameter which is the default kernel for SVM algorithm and noted the results obtained. Later, linear kernel was used as kernel optimization parameter to improve the results obtained by using the rbf kernel.

Stage 5 – Performance Evaluation and Comparison

The results need to be evaluated in order to measure the efficiency of the model. For evaluation we use the different metrics like accuracy, f-score, recall and precision. We compare the results of both the kernels (rbf and linear) using these metrics.

V. RESULTS

This section shows the results of the implementation of the SVM algorithm used for sentiment classification. Table I contains the result of SVM approach with linear as well as rbf kernel. Dataset consisting of Amazon Reviews in a JSON file was used.

A. SVM Linear Kernel

If the data can be linearly split using a single line, then the Linear Kernel is the appropriate choice for processing the data. It is one of the kernels that is employed on a widespread scale. Its primary use occurs in situations where a given data set has a significant number of features. One of the instances that demonstrates a large number of characteristics is text classification; this is because each letter represents a new feature. Therefore, the Linear Kernel approach constitutes the majority of our Text Classification.

B. SVM RBF Kernel

The term "Kernel Function" refers to a procedure that takes data as its input and converts it into the format that is necessary for processing data. The term "kernel" is applied because the Support Vector Machine makes use of a collection of mathematical functions that provide the "window" to alter the data. In most cases, the Kernel Function will adjust the training set of data in such a way that a non-linear decision surface will be able to be transformed into a linear equation in spaces with a greater number of dimensions. In its simplest form, it computes and returns the inner product of the distance between two locations in a standard feature dimension. The transformation was enhanced by the use of the Radial Basis technique.

TABLE I. RESULT ON THE AMAZON DATASET

Method	***Performance Measure***	***33% Test Data***
SVM with Linear Kernel	Accuracy	81.24%
	F-Score	0.812
	Recall	0.812
	Precision	0.812
SVM with rbf Kernel	Accuracy	83.84%
	F-Score	0.764
	Recall	0.838
	Precision	0.703

VI. CONCLUSION

The sentiment analysis of book reviews using a supervised technique is the primary emphasis of this work. I constructed a dataset made out of Amazon Reviews, and applied the widely known support vector machine (SVM) algorithm to it.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 12, December 2022

According to the findings, SVM with rbf kernel produced more accurate results, while SVM with linear kernel produced superior results overall.

REFERENCES

[1] V. Kaur, "Sentimental Analysis of Book Reviews using Unsupervised Semantic Orientation", 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)

[2] P. Walia, V. K. Singh and M. K. Singh, "Evaluating machine learning and unsupervised semantic orientation approaches for sentiment analysis of textual reviews," 2012 IEEE International Conference on Computational Intelligence and Computing Research, 2012, pp. 1-6.

[3] Aljoharah Almjawel, Sahar Bayoumi, Dalal Alshehri, Soroor Alzahrani and Munirah Alotaibi, "Sentiment Analysis and Visualization of Amazon Books' Reviews", 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)

[4] K.S. Srujan; S.S. Nikhil; H. Raghav Rao; K. Karthik; B.S. Harish; H.M. Keerthi Kumar, "Classification of Amazon Book Reviews Based on Sentiment Analysis"

[5] Mounika Addanki and Dr. S. Saraswathi, "Classification of book reviews based on sentiment analysis: A Survey".

[6] J. E. T Akinsola, "Supervised Machine Learning Algorithms: Classification and Comparison", International Journal of Computer Trends and Technology, vol. 48 no. 3, 2017.