iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35291

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

Multi Class Data Classification to Improve Accuracy in Sentiment Analysis using Machine Learning

Daram Vishnu[1], Mothe Vishnu Vardhan Reddy[2], Jagganagari Jaya Prakash Reddy[3], M. K. Jeevan Reddy[4]
[1, 2, 3]*U.G. Student,* [4]*Associate Professor Dept. of ECE, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana-501301 India*

Abstract: ***Sentiment analysis means classifying a text into different emotional classes. These days most of the sentiment analysis techniques divide the text into either binary or ternary classification in this paper we are classifying the movie reviews into 5 classes. Multi class sentiment analysis is a technique which can be used to know the exact sentiment of a review not just polarity of a given textual statement from positive to negative. So that one can know the precise sentiment of a review . Multi class sentiment analysis has always been a challenging task as natural languages are difficult to represent mathematically. The number of features are also generally large which requires huge computational power so to reduce the number of features we will use parts-of-speech tagging using textblob to extract the important features. Sentiment analysis is done using machine learning, where it requires training data and testing data to train a model. Various kinds of models are trained and tested at last one model is selected based on its accuracy and confusion matrix. It is important to analyze the reviews in textual form because large amount of reviews is present all over the web. Analyzing textual reviews can help the firms that are trying to find out the response of their products in the market. In this paper sentiment analysis is demonstrated by analyzing the movie reviews, reviews are taken from IMDB website.***

Keywords: ***Sentiment analysis, parts-of-speech tagging, Textblob, training data, testing data, accuracy and confusion matrix***

I. INTRODUCTION

Sentiment analysis refers to use of NLP, linguistics, to identify the subjective information in the text. It is used to determine the attitude or emotion of the writer with respect to the topic As over the top(OTT) have been popular during this decade, the online content creators and producers ask their audience to share their opinions about the movies and web series they have made. Everyday millions of reviews are generated all over the Internet about different movies, web series . This has made the Internet the most important source of getting ideas and opinions about a movie or a web series. However, as the number of reviews available for a movie grows, it is becoming more difficult for a potential audience to make a good decision on whether to watch the movie or not. Different opinions about the same movie on one hand and ambiguous reviews on the other hand makes audience more confused to get the right decision . Here the need for analyzing this contents seems crucial for all producers and content creators.

Sentiment analysis and classification is a computational study which attempts to address this problem by extracting subjective information from the given texts in natural language, such as opinions and sentiments. Different approaches have used to solve this problem from natural language processing, text analysis, computational linguistics, and biometrics. In recent years, Machine learning methods have got popular in the semantic and review analysis for their simplicity and accuracy.

Amazon prime and Netflix is one of the OTT giants that people are using every day for online binge watching where they can read thousands of reviews dropped by other audience about their desired movies. These reviews provide valuable opinions about a movie such as its intensity, quality and recommendations which helps the audience to understand almost every detail of a movie. This is not only beneficial for audience but also helps content creators who are creating their own movies to understand the audience and their needs better. The overall semantic of audience reviews is determined by determining the sentiment into one of the five classes. In this project, we uses supervised techniques to determine the overall semantic of audience reviews by classifying them into positive and negative sentiment. The data used in this study is a set of movie reviews from IMDB that is collected manually. The five classes in the dataset are

A. 1-negative
B. 2-somewhat negative
C. 3-neutral
D. 4-somewhat positive
E. 5-positive

II. LITERATURE SURVEY

In the paper " Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis" [1] the determined the polarity of the text using boag-of words model. In the paper Sentiment Analysis of Movie Reviews using Machine Learning Techniques[2] they used weka tool to implement binary sentiment classification of movie reviews

Mrs. R.Nithyaet al. [3] represented Sentiment analysis that mainly on subjective and polarity detection. A proposed work include: (i) Feature Extract- Commonly, Sentiment analysis uses machine learning algorithm and a method to extract features from texts and then train the classifier. (ii) Preprocessing- stemming refers reducing words to their roots. Porter's stemming algorithm used for removing stop words. Mostly, adjective words have sentiment. (iii) Product aspects- Textstat is a freely available that can be used for extracting pattern. (iv) Find polarity of opinionated sentence- here SentiStrength lexicon-based classifier used to detect sentiment strength. Here, 575 reviews have been taken from shopping sites.

In the paper sentiment analysis on movie reviews published by Asiri Wijesinghe[4] . This paper uses Tf-idf vectorizer for feature extraction but doesn't use any data filtering techniques. It divides the the sentiment of the reviews into 5 classes

III. PROPOSED SYSTEM

In this project we are going classify the given movie reviews into one of the five sentiment classes. In order to train a model the training data is required, this data is scraped from IMDB website using scrapestorm tool. This data set consists of movie reviews that are took from IMDB website. Supervised machine learning technique is used in this project, where the model is trained by a labelled data set . Before training the model, the reviews that are extracted from IMDB are processed. Here data set consists of reviews and their corresponding ratings. This data is being run on different models like linear classification, logarithmic classification, random forest classifier, support vector machine and naïve bayes algorithm etc. with the same dataset all the models are trained and tested, accuracy of each model is taken into consideration and the model with highest accuracy is selected for analyzing the sentiment class (1,2,3,4,5) in the given reviews.

IV. MACHINE LEARNING METHODS USED

A. Logistic Regression

Logistic regression is a statistical machine learning model that in its basic form uses a logistic function to model a binary variable , although more complex extensions exist. In multivariate analysis , logistic regression (or logit regression) is estimating the parameters of a logistic model (a sort of binary regression).

B. SVM

support-vector machines (SVMs) are supervised learning models with associated learning algorithms that analyse data for classification and multivariate analysis a support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which will be used for classification, regression, or other tasks like outliers detection. Intuitively, a decent separation is achieved by the hyperplane that has the largest distance to the closest training-data point of any class (so-called functional margin), since generally the larger the margin, the lower the generalization error of the classifier

C. Random Forest

Random Forests is the training method that are used for classification and regression problems. It construct's variety of decision trees at training time. To classify new test set it sends the new test set to every of the trees. Each tree perform classification and output a category . The output class is chosen based on majority voting that's the utmost number of comparable class generated by various trees is taken into account because the output of the Random Forest. Random Forests are simple to learn and can be used by both professionals and laypeople with little research and programming required. It can easily be utilized by persons that don't have a robust statistical background

D. K- Nearest Neighbour

K-Nearest neighbours is that the simplest of all the machine learning algorithms. The principle behind this method is to seek out a predefined number of training samples closest in distance to the new point and predict the label from these. the amount of samples are often a user-defined constant or vary based on the local density of points. the distance are often any metric measure. Standard Euclidean distance is that the commonest choice for calculating the space between two points. the closest neighbours are successful during a sizable amount of classification and regression problems, including handwritten digits or satilite image processing

E. Decision Tree

It uses a decision tree (as a predictive model) to travel from observations about an item (represented within the branches) to conclusions about the item's target value (represented within the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that cause those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Decision trees are among the most popular machine learning algorithms given their intelligibility and ease

V. METHODOLOGY

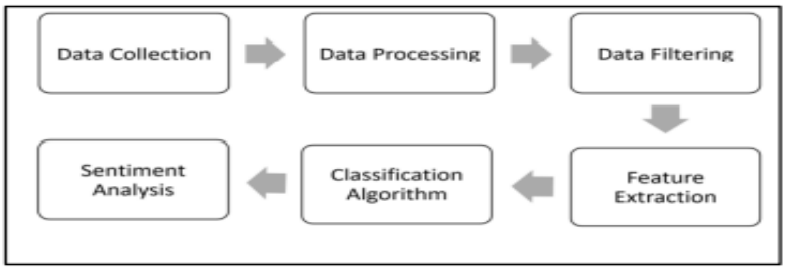


Figure 1:Frame work for Sentiment Analysis

1) *Data Collection:* Data was collected from 16000 user-created movie reviews from IMDB website and this data is collected using web scraping tool Scrapestorm. 12000 of these values are used for training the model 4000 reviews are used as testing dataset.
2) *Data Processing:* Data processing involves Tokenization which is the process of splitting the reviews into individual words called tokens. Tokens are often split using whitespace or punctuation characters. It are often unigram or bigram counting on the classification model used. The bag-ofwords model is one of the widely used model for classification. It is based on the fact of assuming text to be classified as a bag or collection of individual words with no link or interdependence. The simplest way to use this model in our project is by using unigrams as features. It is just a collection of individual tokens or words in the text to be classified, so, we split each review using whitespace. For example, the review “how are you” is split from each whitespace as follows. { “how”, “are”, “you” } The next step in data processing is normalization by conversion of review into lowercase. Reviews are normalized by converting it to lowercase which makes its comparison with an dictionary easier. The following function is used as shown in fig. 2

```
def toLowercase(sentence: Sentence): Sentence =
  sentence.map(_.toLowerCase)
```

Figure 2:Code snippet for converting reviews to lower case

3) *Data Filtering*

a) *Data Filtering:* A review acquired after data processing still has a portion of raw information in it which we may or may not find useful for our application. Thus, these reviews are further filtered by removing stop words, numbers and punctuations.

b) *Stop Words:* for instance , reviews contain stop words which are extremely common words like “is”, “am”, “are” and holds no additional information. These words serve no purpose and this feature is implemented employing a list stored in stopfile.dat. We then compare each word in a tweet with this list and delete the words matching the stop list as shown.In our project we will not be removing stop words as it may sometimes change entire meaning of the sentence

```
import nltk.corpus
nltk.download('stopwords')
from nltk.corpus import stopwords
stop = stopwords.words('english')
df['review'] = df['review'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop)]))
```

Figure 3:Code snippet for removing stop words

c) Removing non-alphabetical characters: Symbols like "#$<" and numbers hold little relevance just in case of sentiment analysis and are removed using pattern matching. Regular expressions are used to identify alphabetical characters only and. The rest of the characters are ignored .This helps to scale back the clutter from the review stream.

d) *Stemming*: It is the process of reducing derived words to their roots. Example includes words like "catch" which has same roots as "catching" and "catches". The library to use stemming is Stanford NLP which also provides various algorithms such as porter stemming. In our case, we have used potter stemmer library for converting words to their base form.

e) *Parts-of speech-tagging:* In this section we will extract the parts of speech for each word in the sentence using TextBlob library

Tags	POS Description
JJ	Adjective
JJR	Adjective Comparative
JJS	Adjective superlative
RB	Adverb
RBR	Adverb comparative
RBS	Adverb superlative
VB	Verb
VBD	Verb Past Tense
VBG	Verb present participle
VBN	Verb past participle
VBP	Verb, known third person singular present
VBZ	Verb third person singular Present.

Figure 4:POS tagging abbrevations

The tags JJ, JJR, JJS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ thses are the tags that are generally determine the sentiment of the text

4) *Feature Extraction:* TF-IDF is a feature vectorization method used in text mining to find the importance of a term to a document in the corpus. TF-IDF converts the sentences in vectors of numbers and this numbers are assigned based on the importance of the word in the documents the to number of features extracted for this data set are 40000

```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf=TfidfVectorizer(use_idf=True, norm='l2',
smooth_idf=True,ngram_range=(1,1))
y=df.sentiment.values
X=tfidf.fit_transform(df['review'].values.astype('U'))
```

Figure 5:Code snippet for TF-IDF transformer

5) *Sentiment Analysis:* The vectors extracted in the feature extraction method are used to train on different machine learning models. After training the data on various machine learning The accuracy of each model is evaluated using the testing dataset. During testing the model each sentence should be cleaned and sentences are converted into vectors. The probability of each class is calculated and the class with highest probability is considered.

A. Software Requirements

1) *Jyputer Notebook:* JupyterNotebook (formerlyIPythonNotebooks) is a web-based interactive computational environment for creating Jupyter notebooks. The "notebook" term can colloquially make reference to many different entities, mainly the Jupyter web application, Jupyter Python web server, or Jupyter document format depending on context. A Jupyter Notebook document is a JSON document, following a versioned schema, containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media, usually ending with the ".ipynb" extension.

A Jupyter Notebook can be converted into a number of open standard output formats like (HTML, presentation slides, LaTeX, PDF, ReStructuredText, Markdown, Python) through "Download As" in the web interface, via the nbconvert library or "jupyter nbconvert" command line interface in a shell. To simplify visualisation of Jupyter notebook documents on the web, the nbconvert library is provided as a service through NbViewer which can take a URL to any publicly available notebook document, convert it to HTML on the fly and display it to the user..

VI. RESULTS

As discussed in the previous segments, we tried various classification models on various feature representations of the textual information in the reviews. Out of these naïve bayes Classifier couldn't even converge for all of our feature sets and hence we could not get a satisfactory answer for it. Among the remaining models, Logistic Regression model seemed to have best performance across all feature representations with classification accuracy around 69%. Also, KNN classifier had the worst accuracy of around 41% across all feature representations. The general order of performance for the model was LogisticRegression > SVM > RandomForestClassifier > Decision tree> kNNClassifier

Model	Accuracy
Logistic regression	69.22
SVM	61
Random forest	55.14
KNN	41.9
Decision tree	52.22

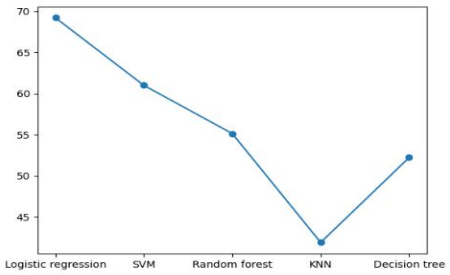


Figure 6: visualization of accuracy results

VII. CONCLUSION

Multi class sentiment classification will give a better insights about the emotion state of the writer compared to binary or ternary classification even tough Sentiment classification is an important tool but it generally requires lot of computational power so it is important to extract important features and parts of speech tagging serves that purpose although it might reduce the accuracy of the model slightly but the number of features required will be reduced so the computational requirement will be reduced.

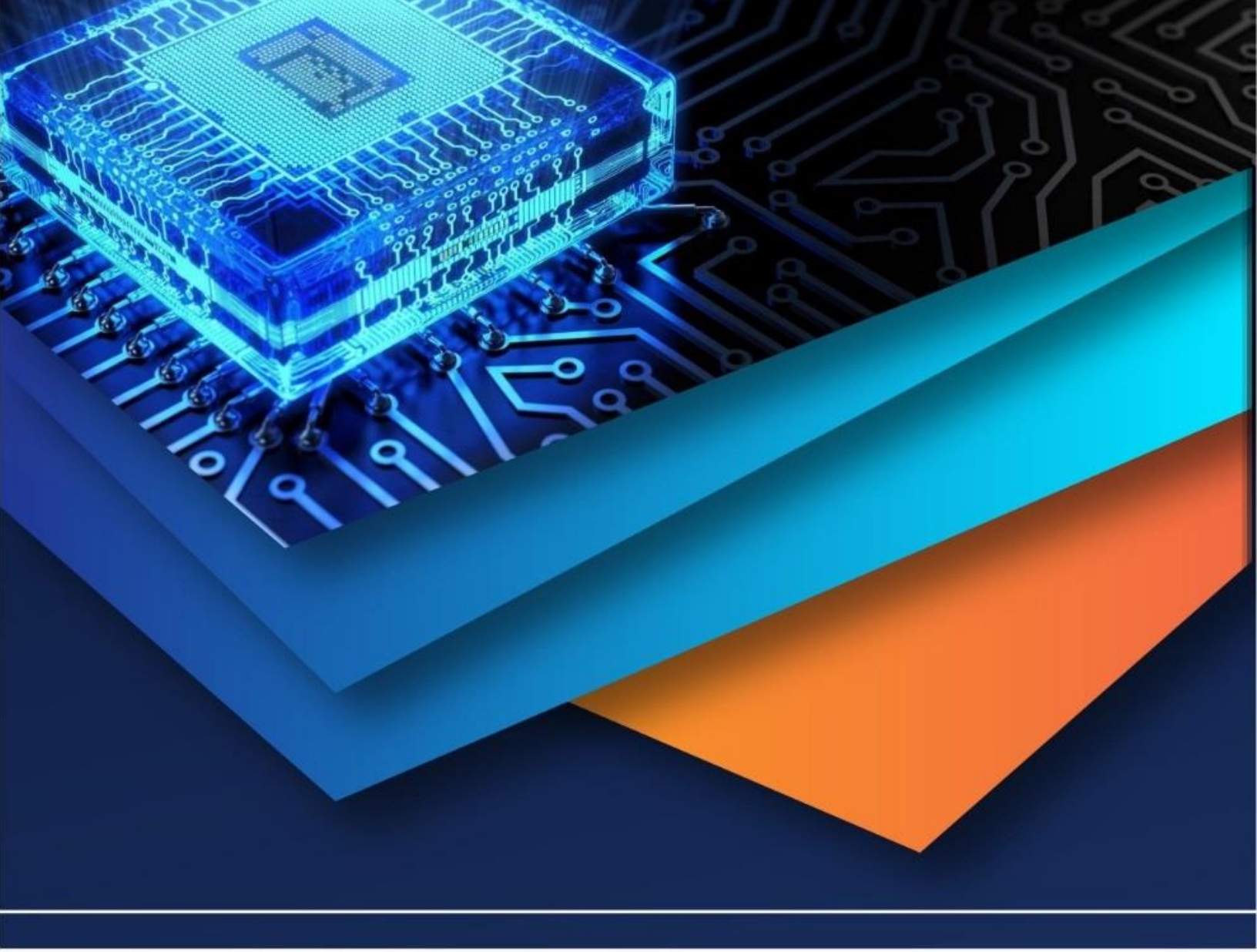
.

VIII. ACKNOWLEDGEMENT

Firstly, we are grateful to Sreenidhi Institute of Science and Technology for giving us the opportunity to work on this project.
We are fortunate to have worked under the supervision of our guide Mr.K.S. Satyanarayana. His guidance and ideas have made this project work.We are thankful to Mr. T. Venkat Rao for being the incharge for this project and conduction reviews.
We are also thankful to the HOD of Electronics and Communication Engineering [ECE], Dr. S.P.V. Subba Rao for giving us access to all the resources that went into buildingthis project

REFERENCES

[1] Doaa Mohey El-Din(2016)" Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis"
[2] Palak Baid , Apoorva Gupta , Neelam Chaplot (2017)Sentiment Analysis of Movie Reviews using Machine Learning Techniques
[3] Mrs. R.Nithya, Dr. D.Maheshwari, "Sentiment Analysis on Unstructured Review", 14 Proceedings of the International Conference on Intelligent Computing Applications IEEE, , pp 367-371, 2014.
[4] https://www.researchgate.net/publication/318532057_Sentiment_Analysis_on_Movie_Reviews



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY