iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** II **Month of publication:** February 2024

DOI: https://doi.org/10.22214/ijraset.2024.58664

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: 7.538
Volume 12 Issue II Feb 2024- Available at www.ijraset.com

Classifying Opinions and Sentiments on Social Networking Sites using Machine Learning Classifiers

Rayees Ahmad Itoo[1], Dr. Yasmin Shaikh[2], Dr. Sanjay Tanwani[3]
[1]*Research Scholar, International Institute of Professional Studies, DAVV, Indore (M. P.), India*
[2]*Assistant Professor, International Institute of Professional Studies, DAVV, Indore (M. P.), India*
[3]*Professor, & Head, School of CS & IT, DAVV, Indore (M. P.), India*

Abstract: ***People now publish evaluations on social media for any product, movie, or location they visit as a result of the Web's rapid development. Customers and product owners can both benefit from the reviews posted on social media in order to assess their offerings. Compared to unstructured data, structured data is simpler to analyze. The reviews are mostly available in an unstructured format. Aspect-Based Sentiment Analysis extracts from the reviews the features of a product and then calculates sentiment for each feature. Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique used to determine the sentiment or emotional tone expressed in a piece of text, such as a sentence, paragraph, or document. Machine leaning classifiers are used to classify sentiments. Machine learning classifiers cannot process raw text so raw text needs to be converted into vectorized form. Feature extraction techniques are used to convert raw text to numerical form also called vectorized data. In present research, four feature extraction techniques with five different machine learning classifiers namely, SVM, Logistic Regression, Naïve Bayes, Random Forest, and KNN are applied to classify sentiments associated with tweets. Two online twitter data sets containing tweets about product reviews and tweets about people's thoughts on public policy are selected for experimentation. In the experiments done, it has been found that the SVM classifier using TF-IDF and HFE shows better performance as compared to other classifiers. Using the feature sets, 97% accuracy and 98% F1-score is achieved in the aspect category prediction task.***

Keywords: Machine Learning, Classifier, Sentiment, Opinions, NLP

I. INTRODUCTION

With the rise of online social media websites like forums, blog sites, reviews and so on, the interest in Opinion Mining (OM) has increased in an exponential manner. Currently, online opinions have become a kind of virtual profit for business companies who are searching for ways to market their products, identify new trends as well as manage their positions [1]. Several organizations utilize OM systems for tracking consumer inputs in online shopping and review sites.

Sentiment analysis is especially useful to companies for analyzing consumer opinions on their items as well as products. Although product features might be clearly stated, finding the basic cause for low profits requires greater focus on individual consumer reviews on those features.

Sentiment Analysis is typically utilized in opinion mining for identifying sentiments, affects, subjectivities as well as other emotional states in online texts. Originally, the task of sentiment analysis was performed on product reviews by processing the products' attributes. However, nowadays sentiment polarity analysis is used in a wide range of domains such as for example the financial domain [2].

II. LITERATURE REVIEW

A. Machine Learning

Machine learning techniques in the classification of sentiment depends on the use of well-known machine learning technology on text data. Supervised learning is an effective classification method and has been used with very promising results for classifying opinions. The regularly used supervised classification techniques in sentiment analysis are Support Vector Machine (SVM), Naïve Bayes (NB) Maximum Entropy (ME), and Artificial Neural Network (NN) and Decision Tree (DT) classifiers. Some less commonly used algorithms are Logistic Regression (LR), K-Nearest Neighbor (KNN), Random Forest (RF),and Bayesian Network (BN).

©IJRASET: All Rights are Reserved

1) Supervised Machine Learning Methods

The machine learning method applied for SA typically belongs to supervised classification as well as text classification methods particularly. In machine learning based classification, two sets of documents are needed: training as well as test sets. The former are utilized by automated classifiers for learning the differentiable features of documents while the latter is utilized for validating the performance of the automated classifier [3]. There are several machine learning methods which have been used for classifying reviews [4].

The goal of machine learning is the development of protocols for optimizing the performance of systems through usage of sample data or previous experiences. Machine learning offers a solution to the classification issue which has two stages. The first stage is to build a learning model from a training dataset. The second stage is to classify the test data on the basis of the trained model. Generally, classification tasks are split into various subtasks:

a) *Data pre-processing:* Previous studies have demonstrated that the preprocessing steps can improve the performance of text retrieval, classification and summarization [5]. Stop words are those words which rarely contribute useful information in terms of document relevance [6]. Tokenization is the most primitive step while processing any document in natural language processing [7]

b) *Feature Selection and /or Feature Reduction:* Feature selection and feature reduction try to decrease dimensionality for the remainder of the steps of the task. Reducing the dimensionality of the data decreases the size of the hypothesis space and there by leads to more rapid implementation time. Generally, features selection methods may be divided into two groups: filters as well as wrappers [8]. Filters are faster than the wrappers and hence are more appropriate for higher dimensional datasets. Diverse features ranking as well as features selection methods are suggested in the machine learning research community like Correlation-based Feature Selection (CFS) [9], Principal Component Analysis (PCA) [10], Gain Ratio (GR) feature evaluation, and Chi-square Feature Evaluation [11], Information gain (IG) [12].

c) *Representation:* In automatic text classification, it has been proved that the term is the best unit for text representation and classification [13]. Though a text document expresses vast range of information, unfortunately, it lacks the imposed structure of traditional database. Therefore, unstructured data, particularly free running text data has to be transformed into a structured data. To do this, many preprocessing techniques are proposed in literature [28, 29].

d) *Classification:* In [16], movie data is reviewed using a range of supervised algorithms such as Naive Bayes, Maximum Entropy, the Stochastic Gradient Descent, and Support Vector Machine. It is proven, however, that the use of unigram, bigram, trigram, and the combination of these, and the mixture of TF-IDF and Counts Vectorizers as the combination to convert texts to a numerical matrix have shown better results of precise classification. Nevertheless, it is also disadvantageous that the Twitter message cannot be checked in limited amounts or in situations where certain statements or symbols reflect the feeling and repetition of the last letter several times. Both these weaknesses can be used to boost the recognition of feelings for potential research.

e) *Post-processing:* Various Pruning techniques, rule filtering, and even knowledge integration are typically included in post processing procedures. These processes all serve as a sort of symbolic filter for the erratic and inaccurate knowledge produced by an inductive algorithm. As a result, the complete chain of data processing should include both pretreatment and post processing procedures. The goal of knowledge discovery research is to provide methods and procedures for sifting through massive databases to find knowledge that is compact, somewhat abstract, but comprehensible, and applicable to other applications. This knowledge is "hidden" in these databases. Knowledge discovery as a nontrivial process of identifying valid, novel, and ultimately understandable knowledge in data [17].

2) Sentiment Classification Methods

Logistic Regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that a given instance belongs to a particular class. In order to diagnose and predict disease, the LR approach is used to predict the outcome of the dependent variable using constant-independent variables [18]. Naïve Bayes (NB) classifier is a simple model for classification. It is simple and works well on text classification. It is a probabilistic classifier on the basis of employing Bayes' theorem with strong independence. NB classifier is a technique that applies to a certain class of problems, namely those that phrased as associating an object with a discrete category. From numerical based approach group, NB has several advantages such as simple, fast and high accuracy [19]. Support Vector Machine: Support Vector Machine (SVM) is a technique for classifying linear data. The fundamental notion behind SVM classification is to discover a maximal margin hyper plane that separates the document vector in a class from the other with maximum margin [20].

SVM is a widely used supervised classifier that has a solid theoretical foundation and performs classification more accurately than most other algorithms in many applications. Many researchers have reported that SVM is perhaps the most accurate method for text classification [21]. It is also widely used in sentiment classification.

III. METHODOLOGY

The steps involved in data collection and preprocessing of twitter data is shown in Figure 1. The data available on Twitter platform contains data on various domains. The first step is to select the data domain. The dataset used for experimentation in this research work is product review data and public policy related data. The data related to selected domain is extracted using Twitter APIs. A review database is prepared with extracted raw tweets. Then the review database is preprocessed by eliminating unwanted duplicate text from raw tweets.

A. Data Selection and Collection

Twitter social media analysis can be used to determine which sentiments are prevalent. This information can be used to make a strategic decision. It also helps to categorize people's feelings and affections. Two online tweet data sets DS-1 (tweets related to product reviews) and DS-2 (tweets related to public opinion on public policy) are selected for experimentation. Firstly the raw tweets related to selected domain are extracted from the most popular networking sites, twitter via Twitter API. Twitter offers three different types of APIs for varied uses. Search APIs are used to extract tweets from a certain user, to extract tweets associated with a given user, or to retrieve tweets related to a custom query based on particular keywords or hash tags.

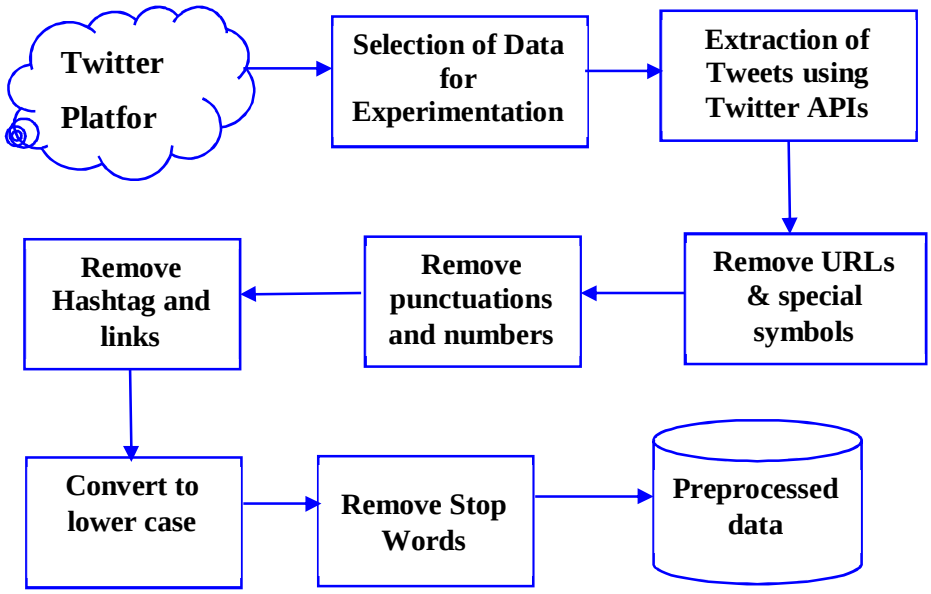


Figure 1: Methodology for Data Collection and Data Preprocessing

A sample of tweets is presented in Figure 2. The extracted tweets are stored in CSV review database along with the user who posted the tweet. A total of 135,000 tweets were retrieved for this study from 13 May 2022 to 16 Sept. 2022, but only 86982 tweets were included for sentiment analysis i.e. 54% of the total tweets. Out of these 33036 are from DS1and 53946 are from DS2. Many duplicate tweets and retweets were found in the retrieved data, and they were removed from the dataset. "Retweets" and duplicates do not present fresh opinions, but they would have altered the overall results.

B. Data Pre-processing

Raw tweets as shown in Figure 2 are to be used to extract the information. The extracted dataset contains information that isn't needed for analysis. Before performing sentiment analysis, unnecessary data from the extracted Twitter dataset, such as hashtags, HTML links, emoticons, punctuations '@', stop words (i.e. is, at the, on), RT, numbers, white spaces, lemmatize, stem Document (i.e. "ruining" will be changed to its root form ruin), and convert tweets to lower case.

	Date	Id	User	Tweet	Location
0	2022-09-16 18:36:38+00:00	1.57E+18	Kevins_Tech	@iHateApplee I have been using Apple products ...	Gettysburg, PA
1	2022-09-15 21:32:02+00:00	1.57E+18	the_echel0n_	@Emre57yyri @sondesix When did I say 1 second?...	USA
2	2022-09-15 17:46:58+00:00	1.57E+18	MAischmann	the picture quality of canon camera is very go...	Diani Beach, Kenya
3	2022-09-15 17:07:41+00:00	1.57E+18	dnaijatechguy	Can never happen. Apple can acquire a smaller ...	Pixelverse
4	2022-09-15 16:43:13+00:00	1.57E+18	eStream_Studios	#Apple #Beats Today's top deals include Paramo...	Gaming and entertainment news
...	...	...	...	...	...
33031	2022-07-20 01:12:55+00:02	8.73E+18	2space	@10810B I remember his commercial for Samsung ...	London
33032	2022-07-13 10:47:55+00:02	1.09E+19	2central_model	@frankdegods @RudaleTheGreat I love apple comp...	Sydney, Australia
33033	2022-07-130 01:12:55+00:03	1.30E+19	7echno	@sparklypadme Haha I love Samsung phones but I...	United States
33034	2022-07-15 10:47:55+00:03	1.51E+19	HandsonToday	@Doog150 A Samsung Galaxy A53, (it was one of ...	UK
33035	2022-07-16 15:30:06+00:00	1.42E+18	Ray_Zinn_	â□□Compared with Samsung and Apple, its averag...	San Jose, CA

33036 rows × 5 columns

Figure 2: Sample Tweets Extracted from Twitter Platform

The extracted datasets databases are then preprocessed using following steps:

1) These URLs and Special characters add no value to text-understanding and cause algorithmic noise. As a result, these are removed to reduce feature space complexity.
2) Punctuation is an important aspect of a sentence since it helps human readers understand it. Punctuation is eliminated from tweets due to the fact that it does not contribute to the training process.
3) Hash tags and link removal are also removed in order to decrease the dataset's complexity.
4) Each character is converted to its lower case using a Python built-in function called conversion to lower case.
5) Stop words are the most frequent terms in a data set. These are deleted to reduce feature space complexity as they add no value to text understanding.

A sample of cleaned dataset after preprocessing is shown in Figure in 3.

	Tweet	Cleaned_Tweets
0	@iHateApplee I have been using Apple products ...	I have been using Apple product...
1	@Emre57yyri @sondesix When did I say 1 second?...	When did I say second ...
2	the picture quality of canon camera is very go...	the picture quality of canon ca...
3	Can never happen. Apple can acquire a smaller ...	Can never happen Apple can ...
4	Today's top deals include Paramount+ and Showt...	Today s top deals include Param...
...	...	...
32979	@10810B I remember his commercial for Samsung ...	I remember his commercial for S...
32980	@frankdegods @RudaleTheGreat I love apple comp...	I love apple computers They ...
32981	@sparklypadme Haha I love Samsung phones but I...	Haha I love Samsung phones but ...
32982	@Doog150 A Samsung Galaxy A53, (it was one of ...	A Samsung Galaxy A ...
32983	â□□Compared with Samsung and Apple, its averag...	Compared with Samsung and...

Figure 3: Sample of Cleaned Tweets

©IJRASET: All Rights are Reserved

C. Feature Extraction Techniques

1) *Bag of Words (BoW):* It constructs a dictionary of set of all words in the given text. It represents all unique words in the form of sparse matrix. Each cell consists of number of times the word occurs ineach text row of document or review. If two vectors are same then they are considered close. Thelimitation of this approach is that it does not take the semantic meaning of words into account.
2) *N grams:* N-grams represent a continuous sequence of N elements from a given set of texts. N grams can be of many types' viz. unigrams, bigrams, trigrams etc.
3) *TF-IDF:* Term Frequency (TF) is the probability of occurrence of a word in the text. The Inverse Document Frequency (IDF) decreases if term frequency is high in reviews. TD-IDF is calculated by multiplying TF with IDF. More importance is given to rare terms and frequent terms in the document. The limitation is that it does not take semantic meaning of words.
4) *Word2Vector (Word2vec):* word2vec place the words in feature space on the basis of their meaning. Words with same meaning are clustered together. It automatically learns the relationship between vectors from raw text. It requires large text corpus for creating vectors and learn relationships among them.
5) *Hierarchical Feature Engineering (HFE):* It is feature space compression space which goes beyond mere feature selection. It utilizes the underlying hierarchical structure of the feature space to generate a much smaller informative feature space for supervised machine learning. Hierarchical feature engineering (HFE) is a combination of feature extraction techniques TF-IDF, BoW and word2vec using feature selection technique chi2.

IV. EXPERIMENTATION

The feature extraction is done using four different feature extraction techniques namely BOW, TF-IDF,Word2vec and HFE. The machine learning models are trained on the extracted features. To evaluate the performance of proposed approach, different classifiers are trained and tested on features extracted from different feature extraction techniques. SVM, Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF) and K-Nearest Nieghbor (KNN) classifiers are trained using BOW, TF-IDF,Word2vec and HFE features. The results of combination of different classifiers with different features are then evaluated to find the effficient combination.

A. Data Splitting

Ideally, the data should be split into three sets: a training set, a test set, and a cross- validation folds or development (dev) set. A brief description of these data sets and the kinds of data they ought to contain is given below:

1) *Training Set:* The training data set is used to build to model. It is a labeled data set that would enable Machine Learning model to learn from this data. For example, a regression model could uncover gradients to lower the cost function by using the instances in this data. These gradients will then be applied to lower costs and improve data prediction.
2) *Cross Validation or Dev Set:* The Cross validation or development set is used to validate the trained model. This isthe most important setting as it will form the basis of model evaluation. If there is a significant disparity between the error on the training set and the error on the development set, the model is over-fitted and has a high variance.
3) *Test Set:* The test set contains the data on which the trained learning model is tested and validated. It tells how efficient the overall model is and how likely is it going to predict class or label for new data item. There are a plethora of evaluation metrics such as precision, recall, accuracy, F1 measure etc. which can be used to measure the performance of the learning model.

For implementing Machine learning model, Python Jupyter Notebook with NLP and machine learning libraries are used. Experiments are conducted using the Scikit, spaCy and NLTK libraries. In order to create a classification model, it is necessary to know how model is actually performing. The tried-and-true technique for doing this is to split the dataset into a training set and a test set. . For this experiment two cleaned and labeled data sets DS-1 and DS-2 described in Section 3.2 are used as input. The data sets are split in the train test ratio as 80:20. Figure 4 illustrates how Machine learning classifier is being trained by using the training data and by using the testing data predicts the labels for inputs

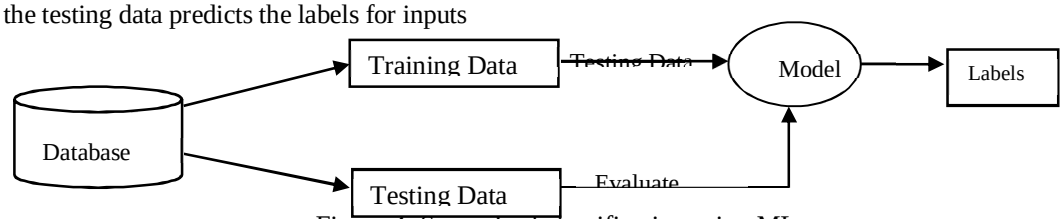


Figure 4: Supervised classification using ML

B. Evaluation Criterion

To evaluate the efficiency of this research, various measures were used, namely: accuracy, precision, recall and F1-Score. These measures have been used in order to evaluate the relevance and irrelevance of the extracted features. *Accuracy* is defined as the ratio of total number of correctly identified aspect polarities to the overall number of aspect polarities. *Precision* P is defined as the ratio of correctly identified terms by the total terms while *Recall* R is defined as the ratio of correctly identified terms by the total identified terms. *F1-Score* is the harmonic mean of precision and recall. F1- score, Precision and recall are directly related. If the value of precision and recall is high, F1-score is high and vice versa.

V. RESULTS AND DISCUSSION

This study uses five Machine learning models Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) and KNN to perform the sentiment analysis aiming to select the one with the best performance. Performanceevaluation of the trained models is realized using accuracy, precision, recall, and F1 score. Results of models are discussed with respect to each featureextraction technique such as BoW, TF-IDF, Word2Vec, and HFE.

A. Results for DS-1

1) Models' Performance using BoW features

The studies are first carried out utilizing BoW characteristics, and the results are shown in Table 1. According to the results, linear and tree-based models significantly outperform other models in terms of precision, recall, and F1 score. By achieving the greatest accuracy, precision, recall, and F1 score of 0.94, 0.93, 0.94, and 0.93, respectively, SVM beats all other models. Performance of RF and KNN is slightly worse,with accuracy scores of 0.85 and 0.77, respectively.

Table 1: Performance of Learning Models Using BoW for DS-1

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.94	0.93	0.94	0.93
Logistic Regression	0.93	0.91	0.92	0.91
Naïve Bayes	0.91	0.92	0.89	0.90
Random Forest	0.85	0.92	0.92	0.92
KNN	0.77	0.93	0.94	0.93

2) Models' Performance using TF-IDF features

Table 2 provides experimental findings for machine learning models utilizing TF-IDF characteristics. According to the results, SVM performs best even when TF-IDF features are employed; it achieves accuracy, precision, recall, and F1 score ratings of 0.96, 0.96, 0.98, and 0.97, respectively. LR comes in second with a 0.95 accuracy score. KNN performs poorly while using TF-IDF features as well. Furthermore, when combined with TF-IDF features, the performance of machine learning models improves. According to these statistics, the model's performance on TF-IDF features is more significant than its performanceon Bow features. In contrast to features created using the TF-IDF method, features created using the BoW technique are primarily simpler and more likely to exist in tweets. However, the TF- IDF feature set is complex because it gives greater weights to rare terms, which are uncommon in tweets. Models thus perform better and have less complexity on BoW features.

Table 2: Performance of Learning Models Using TF-IDF for DS-1

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.97	0.96	0.98	0.97
Logistic Regression	0.95	0.96	0.94	0.95
Naïve Bayes	0.94	0.97	0.92	0.94
Random Forest	0.91	0.96	0.92	0.94
KNN	0.92	0.95	0.91	0.93

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

3) Models' Performance using Word2Vec features

Table 3 displays the effectiveness of models utilizing word2vec features. The results show that when models are trained on Word2Vec characteristics, the performance of Classification is significantly decreased. For instance, the accuracy of the top-performing SVM model is 0.93, compared to 0.94 and 0.96 with BoW and TF-IDF, respectively. Similar to this, LR's accuracy has dropped from 0.94 when used with TF- IDF. However, when combined with Word2Vec characteristics, Naïve Bayes performance has improved, reaching an accuracy score of 0.92 as opposed to 0.91 and 0.90 with BoW and TF-IDF, respectively. As more terms that are connected to one another are employed, Word2Vec's mapping of word embedding into feature space becomes more complicated.

Table 3: Performance of Learning Models Using Word2Vec features for DS-1

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.93	0.94	0.93	0.93
Logistic Regression	0.93	0.94	0.93	0.93
Naïve Bayes	0.92	0.93	0.89	0.91
Random Forest	0.72	0.93	0.89	0.91
KNN	0.71	0.89	0.90	0.89

4) Models' performance using HFE

Table 4 displays the performance of models utilizing the suggested HFE characteristics. The performance of the models is significantly better with the proposed features as compared to BoW, TF-IDF, and Word2Vec. SVM outperforms all models with the best accuracy of 0.97 and scores 0.97 for precision and F1 score and 0.98 for recall. Followed by SVM, the performance of LR is marginally low with a 0.95 accuracy score. The close figures for F1 scores and accuracy show how well these models are working. When combined with HFE features, machine learning models have, for the most part, performed better.

Table 4: Performance of Learning Models Using HFE for DS-1

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.97	0.96	0.98	0.97
Logistic Regression	0.95	0.96	0.94	0.95
Naïve Bayes	0.94	0.94	0.93	0.93
Random Forest	0.92	0.92	0.92	0.92
KNN	0.92	0.93	0.89	0.91

According to the results, NB, RF and KNN who did poorly with BoW, TF-IDF, and Word2Vec, performs significantly better with HFE and achieves an accuracy score of 0.94, 0.92 and 0.92 respectively with favorable values for other performance evaluation measures. In contrast, the proposed HFE provides a set of important features to the machine learning models and produces better results.

B. Visualization of Results for DS-1

The comparative analysis of all machine learning algorithms that are used for performing the task of classifying of text for DS-2 shown in Figure 5.

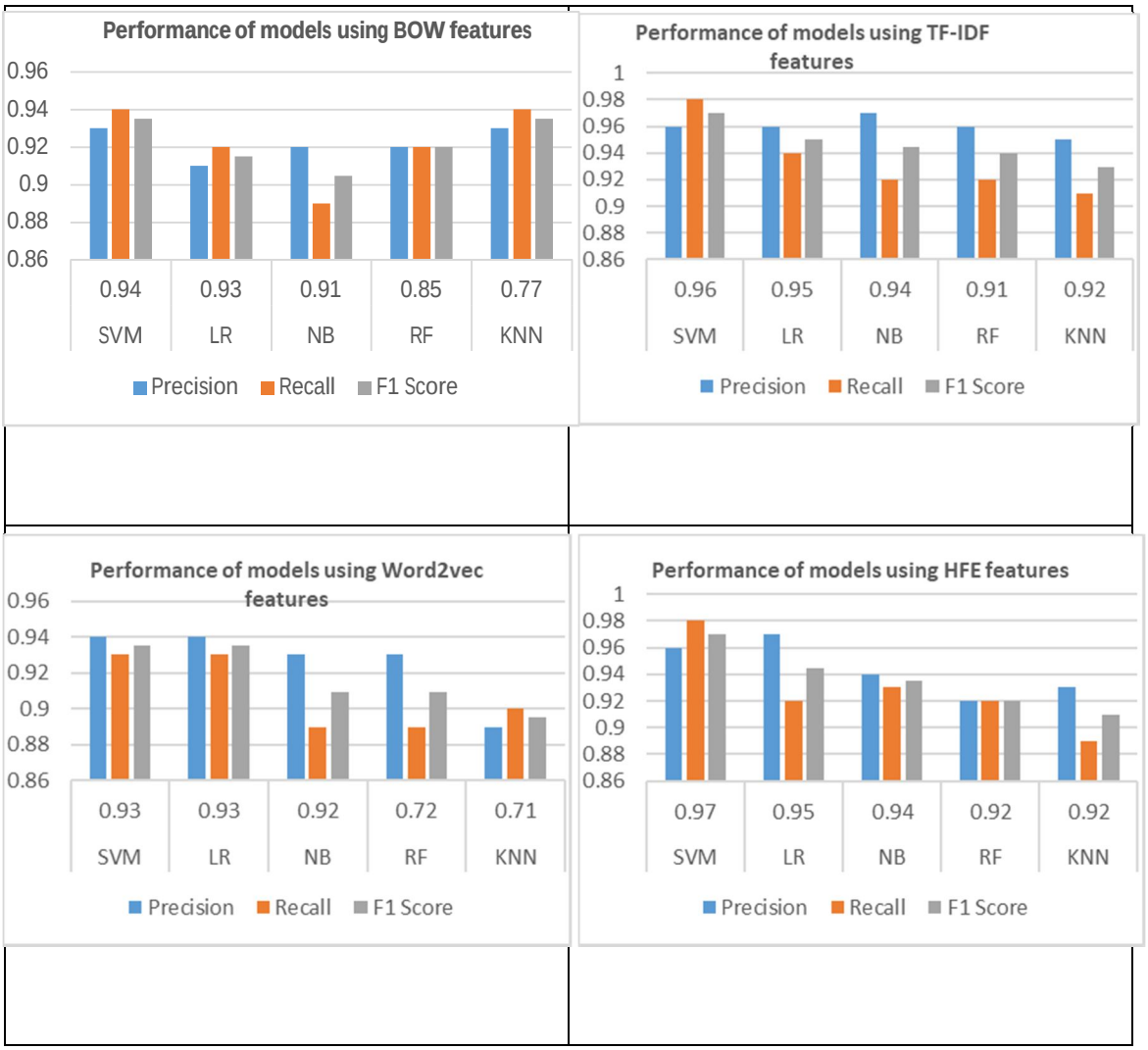


Figure 5: Accuracy Comparison of Classifiers using BoW, TF-IDF, word2vec, and HFE for DS-1

C. Results for DS-2

1) Models' Performance using BoW Features

Table 5 shows the performance results for classifiers utilizing BoW characteristics. According to the results, linear and tree-based models significantly outperform other models in terms of precision, recall, and F1 score. By achieving the greatest accuracy, precision, recall, and F1 score of 0.95, 0.93, 0.94, and 0.94, respectively, SVM beats all other models. Performance of RF and KNN is slightly worse,with accuracy scores of 0.87 and 0.77, respectively.

Table 5: Performance of Learning Models Using BoW for DS-2

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.95	0.93	0.94	0.94
Logistic Regression	0.94	0.92	0.91	0.91
Naïve Bayes	0.91	0.92	0.90	0.90
Random Forest	0.87	0.92	0.92	0.92
KNN	0.77	0.90	0.94	0.93

2) Models' Performance using TF-IDF features

Table 6 provides experimental findings for machine learning models utilizing TF-IDF characteristics.

Table 6: Performance of Learning Models Using TF-IDF for DS-2

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.97	0.97	0.98	0.98
Logistic Regression	0.96	0.95	0.94	0.95
Naïve Bayes	0.94	0.96	0.93	0.94
Random Forest	0.91	0.96	0.92	0.94
KNN	0.92	0.95	0.92	0.93

According to the results, SVM performs best even when TF-IDF features are employed; it achieves accuracy, precision, recall, and F1 score ratings of 0.97, 0.94, 0.91, and 0.92, respectively. LR comes in second with a 0.96 accuracy score. KNN performs poorly while using TF-IDF features as well. Furthermore, when combined with TF-IDF features, the performance of machine learning models improves.

3) Models' Performance using Word2Vec features

Table 7 displays the effectiveness of models utilizing word2vec features. The results show that when models are trained on Word2Vec characteristics, the performance of classification is significantly decreased. For instance, the accuracy of the top-performing SVM model is 0.94, compared to 0.95 and 0.97 with BoW and TF-IDF, respectively. Similar to this, LR's accuracy has dropped from 0.96 when used with TF- IDF. However, when combined with Word2Vec characteristics, Naïve Bayes performance has improved, reaching an accuracy score of 0.92 as opposed to 0.91 and 0.94 with BoW and TF-IDF, respectively. As more terms that are connected to one another are employed, Word2Vec's mapping of word embedding into feature space becomes more complicated.

Table 7: Performance of Learning Models Using Word2Vec features for DS-2

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.94	0.93	0.93	0.93
Logistic Regression	0.93	0.93	0.93	0.93
Naïve Bayes	0.92	0.93	0.91	0.91
Random Forest	0.82	0.93	0.89	0.91
KNN	0.71	0.89	0.90	0.89

4) Models' Performance using HFE

Table 8 Table displays the performance of models utilizing the suggested HFE characteristics. The performance of the models is significantly better with the proposed features as compared to BoW, TF-IDF, and Word2Vec. SVM outperforms all models with the best accuracy of 0.97 and scores 0.98 for precision, 0.96 for recall and 0.97 for F1 score. The Comparative analysis of all machine learning algorithms that are used for performing the task of classifying of text are shown in figure 5.6 (DS2). The results showed that SVM algorithm shows better results than all other algorithms by having 97% Accuracy, precision 98%, recall 96% and F1-score 97% unlike LR, Naïve Bayes, Random forest and KNN also shows good results with accuracy 96%, 94%, 93% and 92% respectively.

Table 8: Performance of Learning Models Using HFE for DS-2

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.97	0.98	0.96	0.97
Logistic Regression	0.94	0.97	0.93	0.94
Naïve Bayes	0.94	0.94	0.93	0.93
Random Forest	0.93	0.92	0.93	0.92
KNN	0.92	0.92	0.89	0.91

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

D. Visualization of Results for DS-2

The comparative analysis of all machine learning algorithms that are used for performing the task of classifying of text for DS-2 shown in Figure 6

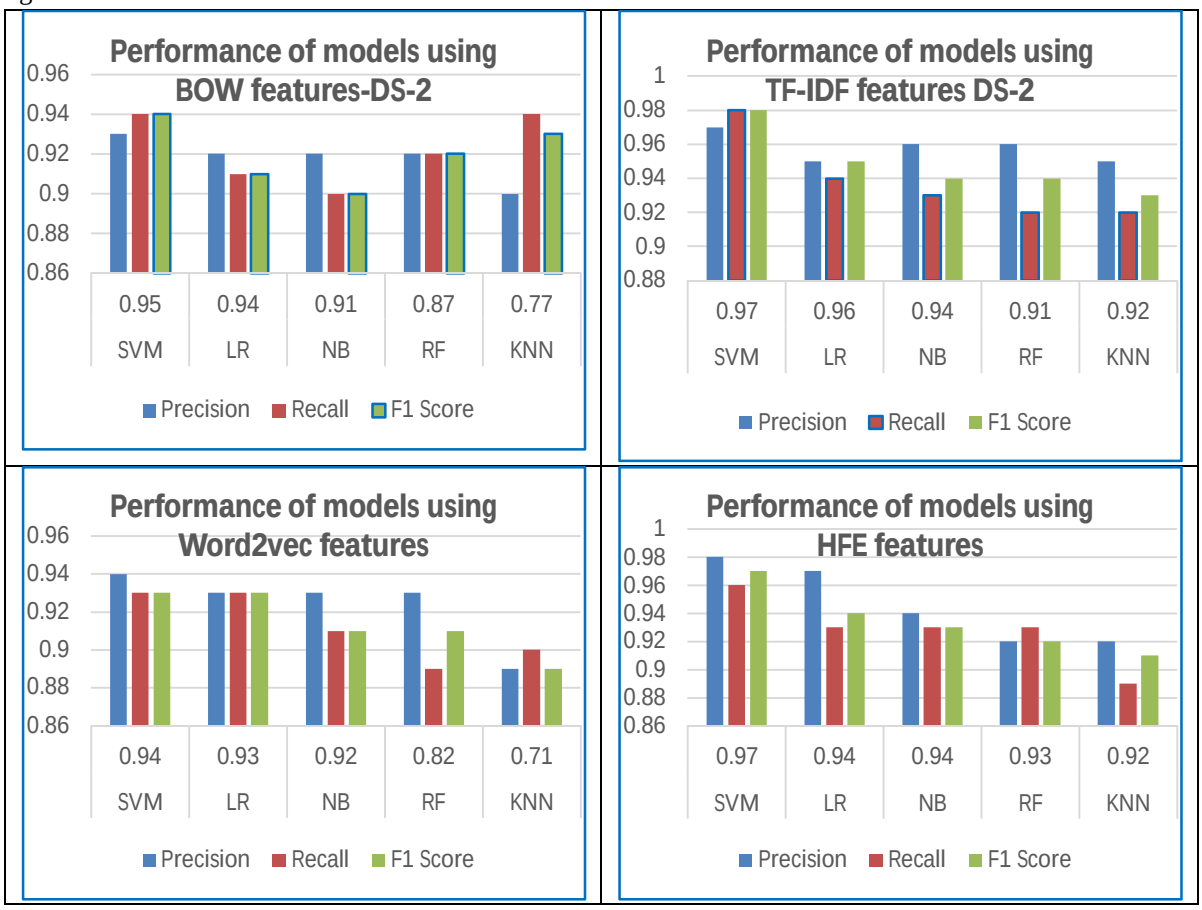


Figure 6: Accuracy Comparison of Classifiers using BoW, TF-IDF, word2vec, and HFE for DS-2

VI. CONCLUSION

In this chapter, two datasets are considered for analysis. The analysis included five popular machine learning classifiers (SVM, Logistic Regression, Naive Bayes, Random Forest and KNN) along with five feature extraction methods (BoW, N-grams, TF-IDF, word2vector, and HFE) to extract sentiments from Twitter dataset. From the experimental study, it has been concluded that the combination of SVM with TF-IDF and SVM with HFE gives best performance for both the datasets.

VII. SUMMARY

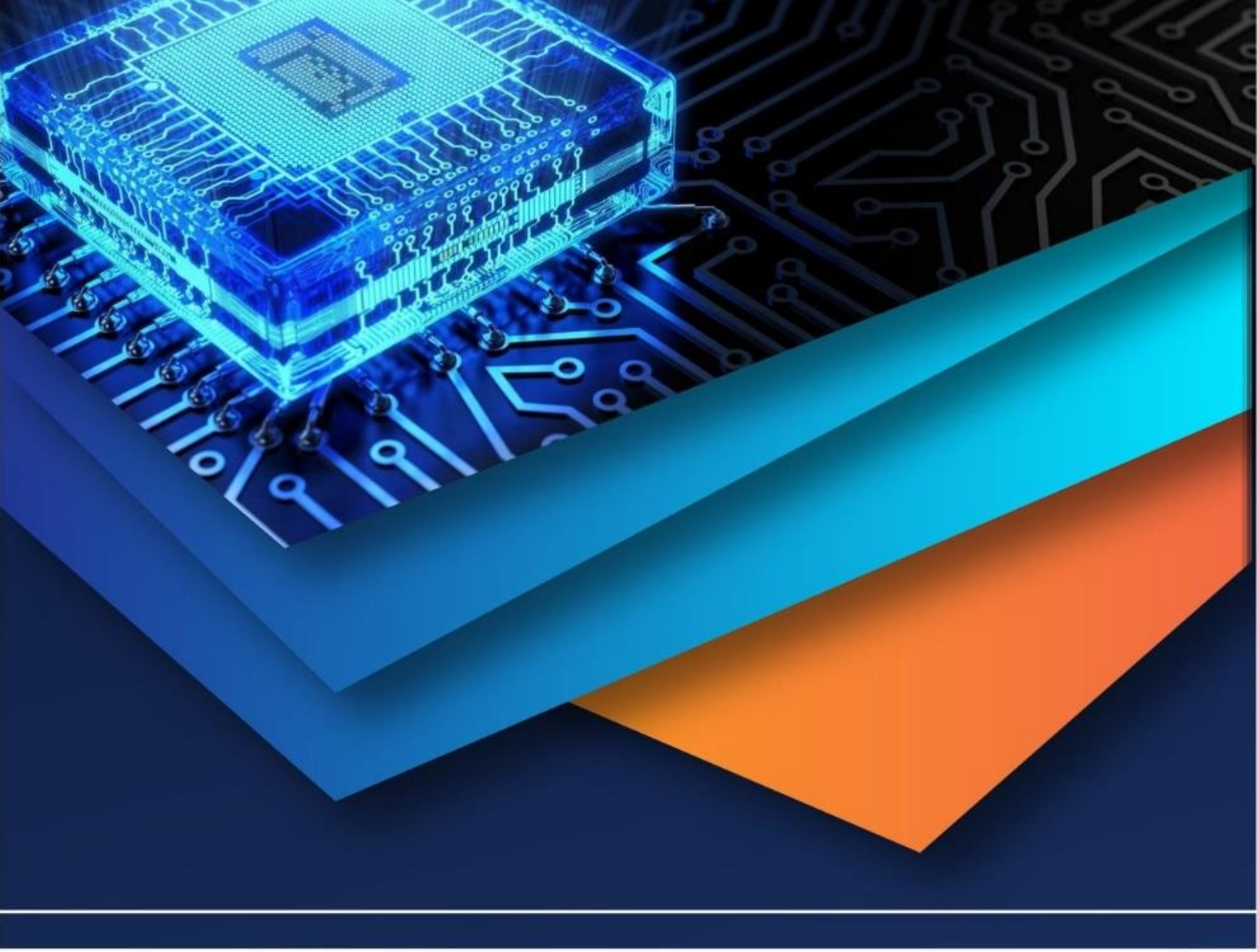
A comprehensive approach to collect and pre-processes unstructured text is presented. The online data is collected from Twitter platform using Twitter APIs. Two different datasets are collected. The raw tweet data is preprocessed and is used for evaluation of proposed approach. A brief description of interconnected activities that are undertaken in order to collect and preprocess raw data is given. The result of preprocessing is cleaned data that can be used for the purpose of data analysis using standard methods.

REFERENCES

[1] Gowtamreddy, P. (2014). Opinion mining of online customer reviews (Doctoral dissertation).
[2] Ruiz-Martínez, J. M., Valencia-García, R., &García-Sánchez, F. (2012, June). Semantic-Based Sentiment analysis in financial news. In Proceedings of the 1st International Workshop on Finance and Economics on the Semantic Web (pp. 38-51).
[3] Vinodhini, G., &Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. International Journal, 2(6).
[4] Ramani, T., &Begam, M. R. (2014). Survey: A Techniques implemented on Opinion Mining. International Journal of Computer Science & Engineering Technology (IJCSET).
[5] Salton G, Singhal A, Mitra M, Buckley C. Automatic text structuring and summarization. Information processing & management. 1997 Mar 1;33(2):193-207.
[6] Tang H, Tan S, Cheng X. A survey on sentiment detection of reviews. Expert Systems with Applications. 2009 Sep 1;36(7):10760-73.

[7] Haddi E, Liu X, Shi Y. The role of text pre-processing in sentiment analysis. Procedia computer science. 2013 Jan 1;17:26-32.
[8] Doraisamy, S., Golzari, S., Mohd, N., Sulaiman, M. N., &Udzir, N. I. (2008, September). A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music. In ISMIR (pp. 331-336).
[9] Karegowda, A. G., Manjunath, A. S., &Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. International Journal of Information Technology and Knowledge Management, 2(2), 271-277.
[10] Eleyan, A., &Demirel, H. (2007). Pca and lda based neural networks for human face recognition. INTECH Open Access Publisher.
[11] Paokanta, P. (2012). β-Thalassemia Knowledge Elicitation Using Data Engineering: PCA, Pearson's Chi Square and Machine Learning. International Journal of Computer Theory and Engineering, 4(5), 702.
[12] Roobaert, D., Karakoulas, G., & Chawla, N. V. (2006). Information gain, correlation and support vector machines. In Feature Extraction (pp. 463-470). Springer Berlin Heidelberg.
[13] Song, F., Liu, S., and Yang, J. 2005. A comparative study on text representation schemes in text categorization,‖ Journal of Pattern Analysis Application, Vol 8, 2005, pp 199 – 209.
[14] Porter, M.F. 1980. An algorithm for suffix stripping. Program, Vol. 14 (3), pp. 130 –137.
[15] Hotho, A., Nürnberger, A., and Paaß, G. 2005. A Brief Survey of Text Mining. Journal for Computational Linguistics and Language Technology. Vol. 20, pp. 19 – 62.
[16] Tripathy, A., Agrawal, A., &Rath, S. (2016). Classification of Sentiment Reviews using N-gram Machine Learning Approach. Expert Systems with Applications, 57(1),117-126 DOI: https://doi.org/10.1016/j.eswa.2016.03.028.
[17] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurasamy, R. Advances in knowledge discovery and data mining. MIT Press, Cambridge, MA, USA, (1996).
[18] Pekkala T, Hall A, Ngandu T, Gils MV, Helisalmi S, Hänninen T, Kemppainen N, Liu Y, Lötjönen J, Paajanen T, Rinne JO. Detecting amyloid positivity in elderly with increased risk of cognitive decline. Frontiers in Aging Neuroscience. 2020 Jul 30;12:228.
[19] Dhande, L. L., &Patnaik, G. K. (2014). Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier.
[20] Gautami, T., &Naganna, S., (2015). Feature selection and classification approach for sentiment analysis. Machine Learning and Applications: An International Journal (MLAIJ), 2(2), 1-16.
[21] Liu B. Opinion mining and sentiment analysis. InWeb data mining 2011 (pp. 459-526). Springer, Berlin, Heidelberg.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)