Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 02 (2026)

Classifying the Small Text into Multiple Sentiment Labels Using Enhaned Features and Machine Learning Technique

Priyanshu Jadon,Dr. Deepshikha Bhatia, Dr. Durgesh Kumar Mishra

Ph.D Scholar, CSE The IIS University, Jaipur,India

Sr. Assistant Professor, CSE The IIS University, Jaipur,India

Director, Symbiosis University of Applied Sciences, Indore, Indi

Abstract:- Social media and e-commerce is become one of the most popular applications. Almost every person has an account on social media and e-commerce. In both platforms, a small size of text is used for expressing opinions by users on a topic or a product. User opinion is one of the sources to understand how a user interacts with a social media post or a product in the e-commerce platform. In this paper, the e-commerce product reviews have been taken into consideration. Product reviews are also playing a vital role in the orientation of the buyer's decision toward product purchasing. In this condition, fake or spam reviews may negatively impact the buyer's decision Thus spam reviews can be a reason for the loss of credibility of the e-commerce platform and also impact buyers. Thus the main aim of the presented work is to identify spam reviews using Machine Learning (ML) technique. The proposed model includes an overview of a modified Term-frequency and inverse document frequency (TF-IDF) to select features from the review text. Then, k-means clustering is used to assign initial class labels to the review text. Finally, for assigning the final classes to the review text a neural network has been used. The experimentation has been conducted on the Amazon product review dataset, which is used for spam review classification tasks. The results were measured and compared, which shows improvement in prediction accuracy as compared to similar spam classification models.

Keywords*: E-commerce, product review classification, small text classification, feature selection of text, limited feature, feature dimensions.*

1. Introduction

Data analysis is an essential part of any business activity. Different areas require data analysis such as marketing, production, consumer feedback, and others. Consumer feedback understanding is an essential aspect of product research and development and the product design department. In this context, the product listed on the e-commerce and the user feedback as the review is also beneficial for improving the product and also increases the sales of the products. But the fake reviews or spam reviews may become a hurdle to understanding the actual feedback about the product.

Therefore it is essential to identify and eliminate the false review of the products. In this work, we are trying to design a machine-learning model, which analyzes the product review data and identifies the spam reviews. The proposed machine learning model utilizes the clustering and neural network technique to provide the classification of reviews. In this chapter, we provide the following vital highlights of the work.

1. Provide a review of recent work based on clustering and opinion mining.
2. Provide a detailed overview of the proposed machine learning scheme.

3. Provide an experimental analysis of the implemented model and comparing them with the relevant technique
4. Provide summary of research work based on observations made during design and experimental analysis. Additionally, future extension plan of the work has also been discussed.

In this section, we provide study about the different techniques and methods which are utilizing the clustering techniques for classifying the social media data or e-commerce platform data.

In energy sector, aim is to providing better service to their customers by knowledge of electricity consumption. Understanding the group behavior of customers is essential. Different shape-based clustering techniques and f-divergence based clustering technique is introduced by ***Y. Zhang et al [1]***. The electricity consumption and customers' consumption patterns are summarized. Squared Hellinger distance and total variation distance are used. The hierarchical clustering is performed using distance matrixes. Using smart meter dataset, this method is compared with other clustering. And consumers' dynamic consumption patterns, relationship between clustering output and other parameters are evaluated. The results demonstrate the given clustering technique is able to create high representative clusters.

Data clustering methods have proven to be capable of extracting useful information from various types and sizes of datasets. ***A. Radovanović et al [2]*** investigates the performance of the agglomerative hierarchical clustering using two time series datasets. The main steps in clustering are presented. Results show that the effectiveness of the clustering algorithm is affected to a large extent by the main characteristics of the clustering data and algorithm's parameters.

A divisive hierarchical clustering algorithm is proposed by ***M. Roux [3].*** It is made in three steps: (1) a splitting procedure for the division of clusters into sub-clusters, (2) an evaluation of the bipartitions resulting and, (3) a formula for determining the node level dendrogram. A set of 12 algorithms is presented and compared. These algorithms are evaluated using the Goodman-Kruskal correlation coefficient. Applied to a hundred random data tables and three real life examples, these comparisons are in favour of methods which are based on unusual ratio-type formulas, namely the Silhouette formula, the Dunn's formula and the Mollineda et al. formula.

Most work on Hierarchical clustering was not considering the optimization, but clustering is an optimization problem, where a "good" clustering is minimizing a cost function. This cost function has certain properties like optimal cost, components are differentiable with hierarchy, and if similarity between elements is identical, then all clustering has same cost. **V. C. Addad et al [4]** define "good" objective functions for clustering techniques. They discuss a set of objective functions. They show performance of the algorithms, are better and faster. They also discuss a beyond worst-case analysis.

Understanding tourism related behavior and travelling patterns is essential for tourism industry. Mostly the segmentation is conducted to develop tourist's profiles for personalizing services. Now days, the data collection problems has solved by wearable sensors in time efficient and less expensive manner. ***J. Rodríguez et al [5]*** described hierarchical clustering for smartphone based geo-localized data. They provide key highlights for the property of consumer mobility. The applicability is demonstrated in the Province of Zeeland. They collected data from 1505 users. The approach resulted in two major clusters and four sub-clusters.

Tourist reviews reflect the tourist's opinions concerning various aspects of a tourist place or service. Extracting aspects from reviews is a challenging task. ***M. Afzaal et al [6]*** propose "enhanced multi-aspect-based opinion classification". First, proposed a probabilistic co-occurrence-based technique, which utilizes the co-occurrence of aspects and sentiment. Next, an aspect extraction technique is introduced, which combine words with aspects for creating the hierarchy. Third, a multi-aspect opinion classification is proposed. The effectiveness of the

model is evaluated. The results demonstrate the supremacy of the given classifiers by delivering 90% accuracy with 87% domain-relevant aspects extraction.

The social media texts have rich information about the complaints, comments, recommendation and suggestion. ***L. Hakim et al [7]*** examines the sentiment from netizensas part of citizen who has vocal sound about the implementation of UU ITE as the first cyber law in Indonesia as a means to identify the current tendency of citizen perception. To perform text mining, Twitter Rest API while R programming was utilized for classification analysis based on hierarchical cluster.

A. A. Tudoran et al [8] has evaluated assembling the existing considerations and key aspects of the online ad-blocking; and exploring the consumers' beliefs and sentiments in terms of expected ad-blocking behavior. Data of 4,093 consumers' in news about ad-blocking has been used. The data are analysed using probabilistic topic modelling and sentiment analysis. Five topics are identified, which is unveiling the structure of consumers' beliefs. A sentiment analysis based on clustered opinions reveals that the opinions focused on the behavior of ads express the negative sentiment. They provide useful insights for practitioners to create/adopt more acceptable ads. It highlights whether ad-free websites have or do not have the potential to become a business opportunity.

S. K. Lakshmanaprabu et al [9] is discussing recommendation by using swarm. The surveys conducted from business locales some features were identified and, fuzzy c-means (FCM) has been used to group these features. The novelty is—the Dragonfly Algorithm (DA)—recognizes features, and a feature-based positioning to discover. The outcomes demonstrate the greatest exactness, 94.56%.

M. Kumar et al [10] sentiment analysis has been performed on movie reviews from BookMyShow. The orientation of the review is interpreted based on the sentiment. The tasks include sentiment analysis with data extraction, clustering and classification. To develop dataset different web pages of movie reviews are taken by an API. The API extract information related to movie name, reviews, rating and other. They present two techniques for categorization. The ROCK is used for clustering and CART for categorizing positive and negative words. Criticisms from user are also extracted. The movie with highest percentage of positive reviews is categorized effectively and accurately.

M. Mittal et al [11] present a method to make easy to predict and analyze health issues by use of tweets. An algorithm has been framed for the same to perform the analysis on health care tweets with association rules to classify the ailments and symptoms through fuzzy set and two step approaches. The results demonstrate the comparison of the WordCloud which concludes that approach of two step authentications the average accuracy of association between the HIV ailments is 98% accuracy and 98% correlation.

People express their opinions on experiences that influence buyers for purchasing products. This data is helpful for analyzing customer preference, needs and behavior. ***S. Riaz et al [12]*** applied sentiment analysis at phrase level on the customer reviews for estimating customer preference by subjective expressions. Then, strength of sentiment word has calculated for indicating expression intensity and applied clustering for organizing words based on intensity. They also compared the results with star-ranking on common dataset and observe result variations. They also offered a visualization of the results for better understanding.

J. S. Deshmukh et al [13] has proposed an approach that extracts and classifies opinion words from a domain called source domain and predicts opinion words of another domain called target domain, which combines modified maximum entropy and bipartite graph clustering. A comparison of classification on reviews of four products is presented. The results demonstrate that method performs well in comparison to the other methods. Comparison of SentiWordNet and domain-independent words reveals that on an average 72.6% and 88.4% words are correctly classified.

D. M. Eler et al [14] analyzing the impact on most mining tasks. They propose a pre-processing method to analyze pre-processing combination to achieve high precision. In order to show combinations of pre-processing, experiments were performed by combinations like stemming, term weighting, term elimination and stop words elimination. These combinations were applied in opinion mining, from which correct classification rates were highlight the strong impact. They provide graphical representations from each combination to show how visual approaches are useful on document similarities and group formation.

P. S. R. Nethravathia et al [15] contributes towards understanding the customer's behaviour dynamics. The methodology includes: (1) customer purchase pattern prediction; (2) augmentation of data set; (3) multiple regression. The analysis shows how the customer hobby has influence on the purchase patterns and fulfilment. They concluded that, by multiple regression, it is possible to evaluate the level of customer satisfaction. The aim is providing guidance in this research using a mix of interdisciplinary methods and techniques.

Objectives

The aim of the proposed work is to apply the ML and sentiment analysis techniques for analyzing business centric data. In this context, a model is presented to accurately classify the opinion of the user towards a product in form of review post. The classification has been done in terms of spam and legitimate review posts. The model for conducting this task is demonstrated in figure 5.1.

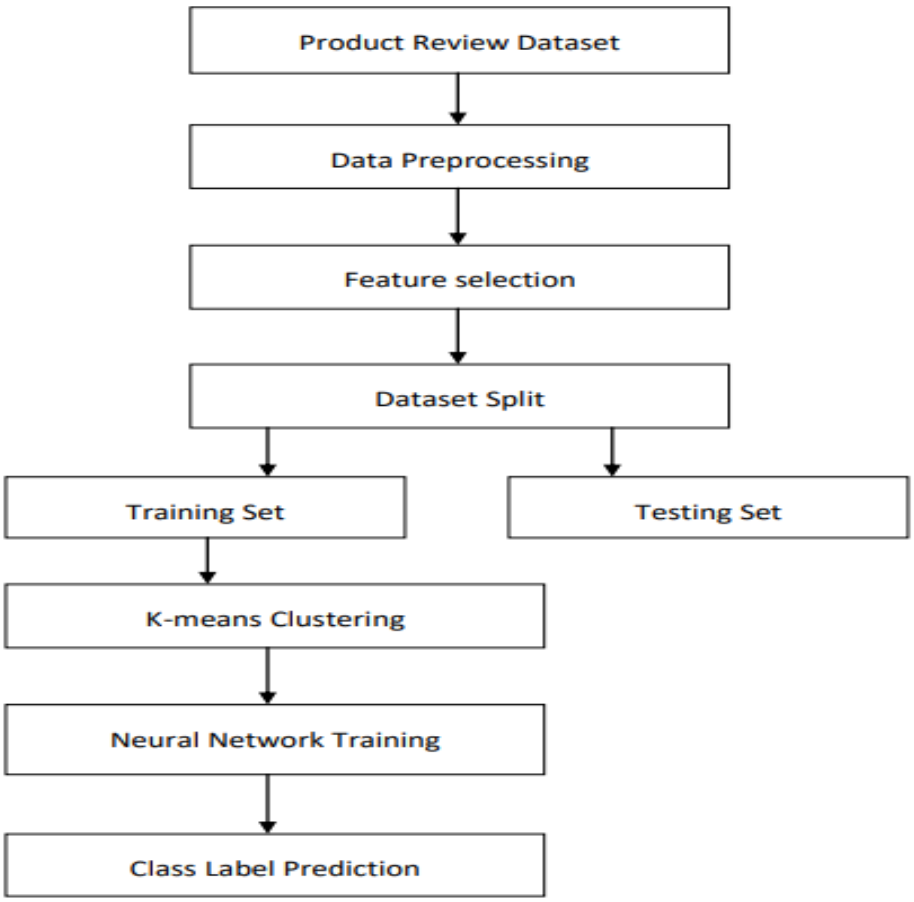


Figure 1 Product review classification model

2. Methods

Product review dataset: the first component of the proposed model is product review dataset. In a ML based system the training samples has been required to deal with leaning problem. Basically the training dataset includes the example patterns which are need to learn. In this presented work the dataset is taken from the Kaggle repository, which is derived from an Amazon product review. The dataset consists of different product categories such as electronics, home and kitchen, toys and games, and many others. Among these product categories the Toys and Games category has been considered for the experiment. The dataset is available in the format of JSON files, which is an XML format. Therefore, first it is needed to parse the JSON files and extract the attributes of the review posts. The attributes of the product review dataset has given table 1.

Table 1 Dataset attributes

S. No.	Attribute Name	Type of data
1	Id	Unique ID of review
2	Reviewer Name	Review name who reviews the product
3	votes-down/up	The ratio of count [down and up]
4	Review Text	Text review post
5	Rating	Rating given
6	Summary	Summary of the review in form of Text
7	Review Time	Time stamp
8	Category	Product category
9	Class	Spam or legitimate

Data pre-processing: The dataset pre-processing is an essential step of ML applications, using the pre-processing techniques the data is transformed in such manner by which the projected data can be utilized with the ML algorithms effectively. The data pre-processing has also used for reducing the noise and improving the quality of information. The refined and high quality data help in better learning. In this presented work, two phases of pre-processing has been used. First the essential attributes from dataset has been identified and in second phase the extracted data has been pre-processed to minimize the noise from dataset.

A. Identifying the potential attributes

The dataset consist of 8 functional attributes and one class label. The attribute “Id” is used for uniquely representing the review instances and it is unique for all the instances. Thus it is less effective for ML algorithm training. Thus, “Id” has been removed initially. Next, attribute is “Reviewer name”, which is an essential attribute to make reviewer profile. Therefore, it keeps with the dataset. Further the “review time” and “category” has been eliminated due to less relevancy of the data analysis context. Therefore after removal of unwanted attributes the dataset have 5 attributes and a class label. The dataset has been next organized into two subsets:

1. **Set 1** contains reviewer name, vote-down/up, rating and class
2. **Set 2** includes a combination of review text and summary

B. Noise reduction from dataset

Both the set of attributes has been treated in different manner. First we consider set 1 which involve a significant amount of missing values therefore the missing values has been filled with the frequent values found in that attributes. Next the set 2 attributes has been considered which consist of textual information. Therefore, the text pre-processing techniques such as special characters and stop words from the input dataset has been removed.

Feature selection: The feature selection is a process which is utilized to identify the significant properties from the given data. The feature selection technique has been considered for set 2 attributes. The set 2 contains the text based data therefore the text features have been extracted. These features are builds with the keywords

which can represent the information domain or subjective knowledge. Additionally, text feature selection is essential because the text is unstructured data type and for utilizing with ML algorithm the transformation of data is essential. In this presented work, we applied two steps of feature extraction form text information:

1. First we identify the essential keywords using the Term frequency and inverse document (TF-IDF) based feature selection technique.
2. Second after constructing the vector using TF-IDF weights the Chi-Square test has been performed between the keywords and the class labels.

Finally based on the chi-square test top 5000 Features are selected for utilization to perform further processing of the text data. The concept of TF-IDF and chi-square test has been given below:

A. TF-IDF

It is a traditional method of text feature selection and representation. The TF-IDF is defined using the following formula:

$$TF = \frac{count\ of\ a\ term\ in\ a\ document}{total\ term\ in\ document}$$

And

$$IDF = \log\left(\frac{N}{df(t)}\right)$$

Where $df(t)$ is Document frequency of a term t, and N is Number of documents containing the term t.

Finally for computing the features in normalized manner we need to calculate the weights for each term using the following formula:

$$w = tf * IDF$$

Using the TF-IDF vector making process, we obtained more than 18000 keywords as feature set. Therefore, in order to reduce the features chi-square test has been performed for finding significant features. Next section provides an over view of chi-square test.

B. Chi-square test

The TF-IDF based generated vector has a significant level of noise. Therefore, to filter irrelevant keywords the Chi-Square Test has been performed. The aim is to identify those keywords which contribute most to the prediction variable. Additionally, avoid Over-fitting and Reduces Training Time. It is to be used when the feature is categorical. It measures the degree of association between two categorical variables. If both are numeric, we can use Pearson's correlation, and if the attribute is numerical and has two classes we can use t-test. Additionally, if more than two classes are available then we can use ANOVA. The Chi-Squared statistics are calculated using the following formula:

$$\varphi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where "O" is actual value and "E" is expected value if two categories are independent. If they are independent then O and E values will be close and if they have some association then the Chi-squared value will be high. By using this test the most significant keywords has been selected and used as vector of 5000 keywords. Before make use of this refined vector the set 1 attributes has also used for computing the score of each review. In this context the review score has been calculated using:

$$Score = \frac{1}{3}\left(\frac{F_r}{T_r} + R + \frac{D}{U}\right)$$

Where, F_r is the frequency of review's review posts, T_r is the total number of reviews, R is the rating given by reviewer, D is the down votes and U is the up votes.

In this scenario, all the factors have been measured in terms of fractions and lies between 0 to 1. Therefore, the compound score is divided by three to get a suitable review score. Finally the feature vector with the use of 5000 keywords, one score variable and a class label has been used to create a final dataset.

Dataset split: The obtained feature vector has a total of 5002 attributes keywords, review score and class label. Additionally a total number of 1943805 dataset instances are available. These final set of data is used for experimentation. Therefore two sets are prepared, first is 75% of entire dataset and used for performing the training and the remaining 25% of the samples are used to test or validate the prepared model.

K-Means Clustering: The prepared training and validation datasets are utilized further for performing training and validation of the ML algorithms. In first step of training we used a clustering algorithm to categorize the data features according to their similarity. For measuring the similarity between two instances of features the Euclidean distance has been used. The Euclidean distance can be expressed as:

$$D(x, y) = \sqrt{(x_i - y_i)^2}$$

Thus the similarity S is expressed as:

$$S = 1 - D(x, y)$$

In order to perform the clustering operation, we utilize the k-means clustering algorithm. The k-means clustering algorithm has worked in the following steps:

Table 2 k-means clustering

Input: Number of categorise k =2, Number of iterations = 50 **Output:** Centroids, Clusters
Process: 1. Select k random instances from training set as centroids 2. Find similarity between selected centroids and feature vectors 3. Based on closeness assign labels to the training vector instances 4. Update centroid by the calculating mean of the current categorize 5. If iterations >= 50 a. Stop 6. Else a. If $\lvert Error_i - Error_{i+1} \rvert \rightarrow 0$ i.Stop b. End if 7. Else a. Iteration +1 b. Go to step 3 8. End if 9. Return centroids, clusters

After training of the k-means clustering algorithm, the algorithm returns k sets of grouped data instances and k optimal centroids. Next by comparing new feature instances with the optimized centroids we assign labels to

each unknown sentence. The predicted class labels have combined with the training set and also with testing dataset. Now after clustering the feature set becomes a total of 5003 attributes.

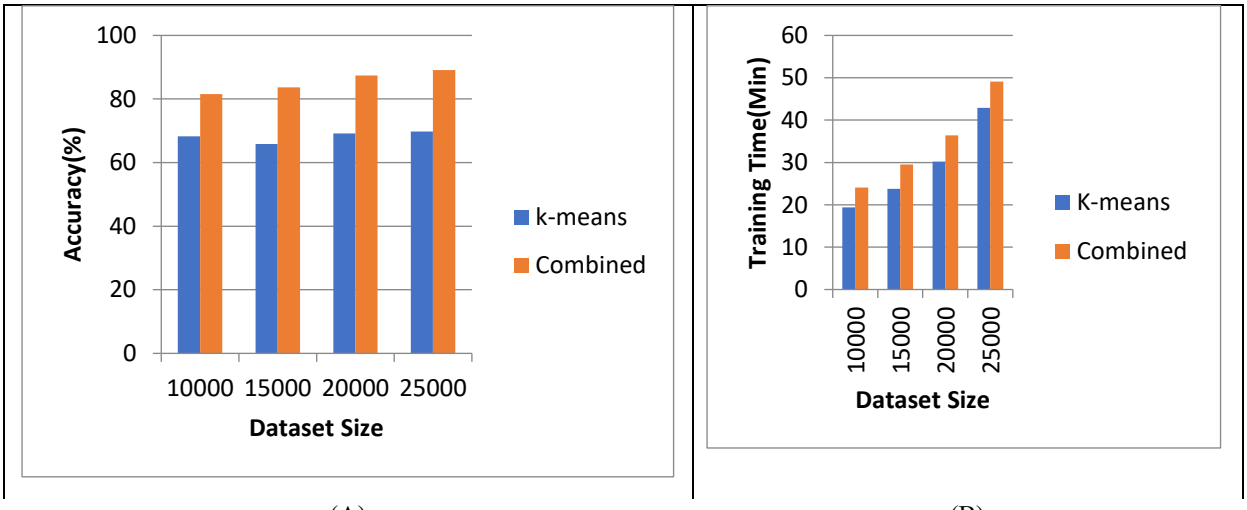
Neural network training: After prediction of initial class label and combination with the previous feature vector a new feature vector has been prepared. Next the deep neural network has been implemented to train and predict the final class labels for the validation set. The configured CNN has been defined as:

1. An input layer is configured with the same dimension of the input neurons as the dataset vector activation additionally the activation function "ReLu" has been used.
2. Next, three dense layers and configured with a total of 128, 64 and 32 neurons. Additionally the activation function "ReLu" has been used.
3. Final layer is also dense layer and configured with two output neurons and "SoftMax" activation function.

Class label prediction: The trained network is finally being used for classifying the test dataset and for validation of the final prediction. Based on the prediction of the test dataset the performance of the classification model has been measured. The obtained performance in terms of classification accuracy and training time is reported into next section.

3. Results

This section provides the experimental results of the proposed spam filtering technique, additionally a comparative study has also conducted between the proposed spam filtering technique and an existing SVM based spam review classification technique.



(A) (B)

Figure 2 Comparing clustering based and proposed model of spam review classification in terms of (A) Accuracy (B) Training Time

A. Accuracy

The proposed technique utilizes two phases of classification first using unsupervised learning technique and second for unsupervised learning technique. Thus accuracy in both the scenarios has been measured and reported in this section. The accuracy describes the total patterns which are accurately defined using the proposed approach. That can be calculated using following equation:

$$accuracy = \frac{total\ correctly\ identified\ spam}{total\ reviews} X100$$

The accuracy of the spam filtering technique with only k-means clustering and combined approach is given in the diagram 2 (A) and table 2. The accuracy is measured here in terms of percentage (%). Additionally the experiments with different dataset size have been conducted.

Table 5.2 Comparing approaches of spam filtering

Data size	Accuracy (%)		Training Time (Sec)		Comparing Accuracy with SVM based model	
	K-means	Combined	K-means	Combined	SVM	Proposed
10000	68.29	81.53	19.4	24.1	78.12	81.53
15000	65.88	83.71	23.8	29.5	75.38	83.71
20000	69.15	87.37	30.2	36.4	79.22	87.37
25000	69.78	89.14	42.9	49.1	79.81	89.14

The X axis shows the dataset size in terms of instances and Y axis shows the accuracy (%). According to the obtained classification accuracy the combined approach which combines the review content analysis and reviewer attributes provide higher accurate results as compared to individual unsupervised learning based classification technique.

B. Training Time

The amount of time required to train a ML algorithm is known as training time. That can be measured using the time difference between algorithm start time and finish time. That is measured using the following formula and in terms of Minute (Min).

$$Training\ time = End\ time - Start\ time$$

The experimental results are obtained based on different dataset size. The training time of conducted experiments has given in figure 2 (B) and table 2. Based on the measured training time, we found that the proposed spam product review classification takes higher time as compared to only k-means based approach. According to the obtained results, we can say the combined approach is time consuming but provides higher degree of accurate results. Additionally as the amount of data increases the amount of training time is also increases.

C. Comparison of the model

As a justification tool the accuracy matrix has been a popular way of comparing the ML algorithms. In this section, a comparison of the proposed spam filtering technique and an available spam review classification technique based on SVM classifier has been performed. The SVM based spam filter technique only considers the TF-IDF based feature extraction. On the other hand, the proposed technique involves an improved TF-IDF based feature selection technique. Additionally utilizes two phases of classification for ensuring the accurate classification. In first phase, the review text is categorized based on a k-means clustering technique and then utilized a supervised learning classifier to learn on additional review parameters. The same dataset is used for validation of both the ML models and results are reported in figure 3 and table 2.

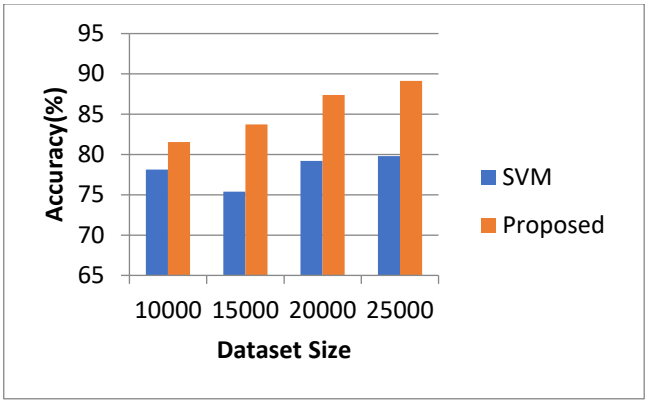


Figure 3 Comparison of the model

Based on the comparative results of the proposed and traditional spam review classification technique is given in figure 3. According to the results the proposed spam filtering technique provides better and consistent accuracy as compared to traditional SVM based spam classification technique. That indicates the strength of including of reviewer's attributes and review content for enhancing the accuracy of the spam review classification. In addition, the accuracy of proposed spam filtering technique is also improved by selecting the suitable keywords as feature.

4. Discussion

In this work, the text analysis has been performed for identifying the spam reviews in e-commerce platforms. It is an essential task because most of the e-commerce consumers are utilizing the reviews to decide the product adoptability. The spam reviews of a product have significant influence on the buyer's decision making process. Therefore, the identification and removal of the false and biased review is an essential work. In this context, the proposed work is focused on developing an improved spam review classification system. The proposed work includes the employment of improved text feature selection technique. Additionally utilization of reviewer's credibility to accurately classify the spam reviews. Therefore, first the raw dataset has been pre-processed to make clean the learning data. Further the dataset has been subdivided into two parts first set of attribute describe the reviewer's profile. Additionally second part of attribute consists of text.

The text based attributes first utilized with the TF-IDF based feature extraction technique. The TF-IDF technique returns the suitable features, but the dimensions of the data is higher. Therefore to reduce the search space the chi-square technique has been applied. Next the reviewer's profile attributes has been used to calculate the review score. The review score and finally refined text features have been combined to prepare a new dataset features. Further the combined features have categorized using the K-means clustering algorithm. After clustering of the dataset features, the k-means based assigned labels are also included into the dataset features. Finally, a neural network is trained on the combined dataset features and the classification of validation has been performed. Using the validation process the results are measured. The Amazon review dataset is considered. Additionally, the proposed model is compared with an existing SVM based spam review classification model. This model utilizes the review text for performing the classification task. Based on the comparative results the proposed method has found more accurate and efficient as compared to the SVM based spam review classification approach.

Refrences

[1] Y. Zhang, Q. Ai, Z. Li, "Grouping of dynamic electricity consumption behaviour: An f-divergence based hierarchical clustering model", IET Generation, Transmission & Distribution, 15, 3164–3175, 2021

[2] A. Radovanović, J. Li, J. V. Milanović, N. Milosavljević, R. Storchi, "Application of Agglomerative Hierarchical Clustering for Clustering of Time Series Data", Proceedings of IEEE PES Innovative Smart Grid Technologies Europe, ISGT-Europe pp. 640-644, 2020

[3] M. Roux, "A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms", Journal of Classification, Vol. 35, No. 2, 345-366, 2018
[4] V. C. Addad, V. Kanade, F. M. Trenn, C. Mathieu, "Hierarchical Clustering: Objective Functions and Algorithms", Journal of the ACM, Vol. 66, No. 4, Article 26. June 2019
[5] J. Rodríguez, I. Semanjski, S. Gautama, N. V. deWeghe, D. Ochoa, "Unsupervised Hierarchical Clustering Approach for Tourism Market Segmentation Based on Crowdsourced Mobile Phone Data", Sensors, 18, 2972, 2018
[6] M. Afzaal, M. Usman, A. C. M. Fong, S. Fong, "Multiaspect-based opinion classification model for tourist reviews", Expert Systems; 36, e12371, 2019
[7] L. Hakim, T. F. Kusumasari, M. Lubis, "Text Mining of UU-ITE Implementation in Indonesia", IOP Conf. Series: Journal of Physics: Conf. Series, 1007, 012038, 2018
[8] A. A. Tudoran, "Why do internet consumers block ads? New evidence from consumer opinion mining and sentiment analysis", Emerald Publishing Limited, Internet Research, Vol. 29, No. 1, pp. 144-166, 2019
[9] S. K. Lakshmanaprabu, K. Shankar, D. Gupta, A. Khanna, J. J. P. C. Rodrigues, P. R. Pinheiro, V. H. C. de Albuquerque, "Ranking Analysis for Online Customer Reviews of Products Using Opinion Mining with Clustering", Hindawi Complexity, Article ID 3569351, 9 pages, Volume 2018
[10] M. Kumar, N. Kumar, "Sentiment Analysis Using Robust Hierarchical Clustering Algorithm for Opinion Mining On Movie Reviews-Based Applications", International Journal of Innovative Technology and Exploring Engineering, 2278-3075, Volume-8, Issue-8, June, 2019
[11] M. Mittal, I. Kaur, S. C. Pandey, A. Verma, L. M. Goyal, "Opinion Mining for the Tweets in Healthcare Sector using Fuzzy Association Rule", EAI Endorsed Transactions on Pervasive Health and Technology, 10, Volume 4, Issue 16, e2, 2018
[12] S. Riaz, M. Fatima, M. Kamran, M. W. Nisar, "Opinion mining on large scale data using sentiment analysis and k-means clustering", Cluster Computing, Springer Science+Business Media, LLC 2017
[13] J. S. Deshmukh, A. K. Tripathy, "Entropy based classifier for cross-domain opinion mining", Applied Computing and Informatics, 14, 55–64, 2018
[14] D. M. Eler, D. Grosa, I. Pola, R. Garcia, R. Correia, J. Teixeira, "Analysis of Document Pre-Processing Effects in Text and Opinion Mining", Information, 9, 100, 2018
[15] P. S. R. Nethravathia, G. V. Baib, C. Spulbarc , M. Suhand, R. Biraue, T. Calugaruc, I. T. Hawaldarf, A. Ejazg, "Business intelligence appraisal based on customer behaviour profile by using hobby based opinion mining in India: a case study", Economic Research-Ekonomska Istrazivanja, VOL. 33, NO. 1, 1889–1908, 2020