IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 15, 2020, accepted March 1, 2020, date of publication March 9, 2020, date of current version March 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2979226

Spam Review Detection Using the Linguistic and Spammer Behavioral Methods

NAVEED HUSSAIN[1], HAMID TURAB MIRZA[1], IBRAR HUSSAIN[2], FAIZA IQBAL[2], AND IMRAN MEMON[3]

[1]Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Lahore 54000, Pakistan
[2]Department of Software Engineering, The University of Lahore, Lahore 54000, Pakistan
[3]Department of Computer Science, Bahira University, Karachi Campus, Karachi 75260, Pakistan

Corresponding author: Hamid Turab Mirza (drturab@cuilahore.edu.pk)

ABSTRACT Online reviews regarding different products or services have become the main source to determine public opinions. Consequently, manufacturers and sellers are extremely concerned with customer reviews as these have a direct impact on their businesses. Unfortunately, to gain profits or fame, spam reviews are written to promote or demote targeted products or services. This practice is known as review spamming. In recent years, the spam review detection problem has gained much attention from communities and researchers, but still there is a need to perform experiments on real-world large-scale review datasets. This can help to analyze the impact of widespread opinion spam in online reviews. In this work, two different spam review detection methods have been proposed: (1) Spam Review Detection using Behavioral Method (SRD-BM) utilizes thirteen different spammer's behavioral features to calculate the review spam score which is then used to identify spammers and spam reviews, and (2) Spam Review Detection using Linguistic Method (SRD-LM) works on the content of the reviews and utilizes transformation, feature selection and classification to identify the spam reviews. Experimental evaluations are conducted on a real-world Amazon review dataset which analyze 26.7 million reviews and 15.4 million reviewers. The evaluations show that both proposed models have significantly improved the detection process of spam reviews. Specifically, SRD-BM achieved 93.1% accuracy whereas SRD-LM achieved 88.5% accuracy in spam review detection. Comparatively, SRD-BM achieved better accuracy because it works on utilizing rich set of spammers behavioral features of review dataset which provides in-depth analysis of spammer behaviour. Moreover, both proposed models outperformed existing approaches when compared in terms of accurate identification of spam reviews. To the best of our knowledge, this is the first study of its kind which uses large-scale review dataset to analyze different spammers' behavioral features and linguistic method utilizing different available classifiers.

INDEX TERMS Online product reviews, spam reviews, spam review detection, linguistic features, spammer behavioral features.

I. INTRODUCTION

Nowadays, the World Wide Web (WWW) has become the main source for individuals to express themselves. People can easily share their views about any product or service by using e-commerce sites, forums and blogs. Everybody on the web is now acknowledging the importance of these online reviews for both customers and vendors. Most people read reviews about products and services before buying them. Vendors can also design their future production or marketing strategies based on these reviews [1]. For example, if various customers buying a specific model of a laptop, post reviews about issues related to its screen design, the manufacturer can be aware and resolve this issue to increase customer satisfaction [2].

Recently, the trend of spam review attacks has increased because anybody can simply write spam reviews and post them online without any constraint. Anyone can hire people to write fake reviews for their products and services, such people are called spammers. Spam reviews are usually written to gain profits or to promote a product or service. This practice is known as review spamming [3], [4]. The main problem with opinion sharing websites is that spammers can easily create hype about the product by writing spam reviews. These spam reviews can play a key role in increasing the value of a product or service [5]. For example, if a customer wants to purchase a product online, he/she usually goes to the review

The associate editor coordinating the review of this manuscript and approving it for publication was Bin Liu.

section to know about other buyers' feedback. If the reviews are mostly positive, the user may purchase it, otherwise, he/she would not buy that specific product [6]. This all shows that spam reviews have become the main problem in online shopping, which can cause loss to both the customers and manufacturers.

Review spam can financially affect businesses and can cause a sense of mistrust in the public; therefore, due to its significance, this problem has recently attracted the consideration of media as well as governments. Recent media news from the New York Times and BBC [7] have stated that ''nowadays, spam reviews are becoming very common on the websites and, recently, a photography company was exposed to thousands of fake consumer reviews''. Hence, detection of spam reviews is critical and without solving this important issue, online review sites could become a place full of lies and, as such, completely useless. To counter this issue, major commercial review hosting sites, such as Yelp[1] and Amazon,[2] have already made some progress in detecting spam reviews [8]. In the last few years, researchers have studied the spam review problem and proposed different techniques. However, there is still a lot of room for improvement in spam review detection techniques using real-world datasets [9], [10].

Review spam is usually related to email and web spam. The web spam is used to attract people by manipulating the content of the page so that the web page will be ranked highly by the search engines [11], [12]. Email spam is mainly used for advertising purposes. However, spam reviews are different in a sense as these give the wrong opinion about a product/service and it is very difficult to detect spam reviews manually. Therefore, existing web spam or email spam detection techniques [13] are not suitable for spam review detection. Spam review detection is a challenging task as no one can detect a review as spam by simply reading its text. Review websites are usually open to public reviews. Therefore, any user can act as spammer to write spam reviews about any product and/or service. Spam reviews appear as legitimate until different spammer behavioral features and/or the review text is analyzed to identify the spam reviews. Based on these perspectives, existing approaches of Spam Review Detection (SRD) utilizes spammer behavioral features or linguistic features for the detection of spammers and spam reviews respectively. The linguistic feature considers review text to identify the reviews as spam or not spam; whereas behavioral features reflect the behavior of reviewer in terms of time stamp of review, review rating, user profile, etc.

From the literature review, it has been observed that existing approaches either adapted linguistic methods or utilized behavioral characteristics separately to identify the spammers and spam reviews. Most of the existing works have only utilized the uni-gram linguistic approach to classify reviews [9]. Usually, the uni-gram approach produces good results but fails in some cases. For example, in the following review; ''This hotel is not good'' when analyzed through the uni-gram approach, gives the popularity of the review as neutral with one positive word ''good'' and one negative word ''not''. But when the same review is analyzed using a bi-gram approach, it gives a negative impression due to the use of the words ''not good''. Considering this limitation, this research intends to utilize N-gram approach to accurately analyze spam reviews. Similarly, many existing approaches ignored several important behavioral features while developing behavioral models for spammer detection. Therefore, there is still a need to employ all existing behavioral and linguistic features to accurately filter out spam and not-spam reviews. The aim of this work is to develop an SRD model adapting a vast set of behavioral and linguistic features on large-scale real-world dataset.

In this study, the investigation about the spam review is based on 26.7 million reviews and 15.4 million reviewers from Amazon.com. However, the main limitation of this domain is that the available datasets are unlabelled, the same is the case with Amazon dataset. To tackle this problem, the proposed approach first formulates a procedure of Spam Review Detection using Behavioral Methods (SRD-BM) to create a labelled dataset. This labelled dataset, then, utilizes Spam Review Detection using Linguistic Method (SRD-LM) to train the classifiers. Specifically, the proposed approaches incorporated linguistics features, such as N-gram techniques, and a number of spammer behavioral features, such as activity window, review count, the ratio of a positive review, the ratio of negative reviews, the ratio of the first review and the review length, for developing the spam review detection model. These behavioral and linguistic features were not properly utilized in previous studies.

This work has made the following research contributions:

1. Proposed methods utilized real-world large-scale Amazon review dataset
2. Proposed SRD-BM which incorporated thirteen different behavioral features to identify spammers and spam reviews
3. Proposed SRD-LM which utilized linguistic features and classifiers to identify spam reviews
4. Compared and analyzed the accuracy of proposed SRD-BM and SRD-LM

The rest of the paper is organized as follows. Section II presents the literature review. Section III describes the statistics of the Amazon dataset. Section IV elaborates the proposed Spam Review Detection using the spammer Behavioral Features Method (SRD-BM). Section V presents the proposed Spam Review Detection using Linguistic Method (SRD-LM). Section VI describes the comparative analysis of SRD-BM and SRD-LM in terms of accurately identifying the spam reviews. Finally, Section VII concludes the work.

II. LITERATURE REVIEW

Existing studies have explored a variety of different spam review detection methods to detect spam reviews. This study

[1] www.yelp.com
[2] www.amazon.com

has reviewed the literature from two perspectives: (1) SRD using the spammer behavioral method and (2) SRD using the linguistic method. The aim is to determine the novel contributions of the proposed work by comparing it with prior studies.

A. SPAM REVIEW DETECTION (SRD) USING THE SPAMMER BEHAVIORAL METHOD

Spam review detection using the spammer behavioral method finds the unusual spammer patterns and relationships between different spammers. Only a few studies have explored spam review detection using the spammer behavioral method to date. For example, Mukherjee *et al.* [14] developed a spam review detection method using a clustering technique by modelling the spamicity of the reviewer to identify spammer and not-spammer clusters. Heydari *et al.* [15] have proposed a model incorporating only time series feature of the reviewer on an Amazon real dataset.

Kc and Mukherjee [16] offered a text mining model by using the unsupervised approach and features, relying upon the time integration among multiple time durations. In addition, this model was integrated with the semantic language model for spotting spam reviews and used a Yelp dataset.[3] Li *et al.* [17] have suggested that the author spamicity unsupervised model has been based on features such as the review posting rate and temporal pattern. The model produced two clusters: spammers and truthful users. The datasets were gathered from the Chinese website Dianping[4] to train the proposed model. Dematis *et al.* [7] have observed a network model for spam review detection. In their work, the correlation among users and products was captured and the algorithm was used to recognize the spam reviews.

Based on the review of spammer behavioral models, it has been observed that most of existing studies [14]–[17] have only utilized time series-based spammer behavioral feature. It is analyzed that utilizing rich set of behavioral features can help in improving the accuracy of spammer identification. Therefore, the proposed behavioral framework utilizes thirteen spammer behavioral features to calculate spam score in spam review identification.

B. SPAM REVIEW DETECTION (SRD) USING THE LINGUISTIC METHOD

The spam review detection problem was first studied in 2007 by Jindal and Liu [18]. They analyzed 5.8 million reviews from Amazon.com. The key focus of this research was on review text. The authors have found many duplications of review content and analyzed that a spammer mostly copies the review content for a different purpose after a little modification. The authors trained the model by using the logistic regression classifier.

Lau *et al.* [19] have applied the semantic language model to identify spam reviews. The authors used the Support Vector Machine classifier to train the proposed method. Li *et al.* [20] used a supervised learning approach with a co-training method to highlight spammers based on linguistic features. Fusilier *et al.* [21] proposed a classification method that used N-gram characters as a linguistic feature. Moreover, the proposed method used the Naïve Bayes to classify spam and not-spam reviews. Ott *et al.* [22] have designed a dataset for spam review detection, employing a crowd source through AMT (Amazon Mechanical Turk). The authors found that the classifier performed better by adding elements such as psycholinguistic features.

Hazim *et al.* [23] used statistically based features for the Extreme Gradient Boost Model and Generalized Boosted Regression Model to evaluate multilingual datasets (i.e., the Malay and English languages). It was observed by the experimental results that the Extreme Gradient Boost Model performed better for the English review dataset and the Generalized Boosted Regression Model performed better for the Malay dataset. Kumar *et al.* [24] have proposed a hierarchical supervised-learning method. This method analyzed reviewer's behavioral features and their interactions using multivariate distribution. Zhang *et al.* [25] recommended a supervised model based on reviewer features to identify spam reviews.

Ahmed and Danti [26] used various rule-based machine learning algorithms. Moreover, the authors compared the effectiveness of the proposed method through a Ten-Fold cross-validation training model for sentiment classification. Lin *et al.* [27] performed different experiments using the threshold-based method to identify spam reviews. The authors proposed different time-sensitive features to find spam reviews as early as possible and trained the model by using the SVM classifier. Li *et al.* [28] used the feature-based sparse additive generative model and the SVM classifier to discover the general rule for spam review detection.

Based on the literature review, it has been observed that most of the existing studies [21], [27], [28] did not incorporate a number of important linguistic features while designing linguistic feature-based SRD models and utilized only one classifier to train their proposed models. The current study, therefore, extends the SRD domain to design a linguistic model utilizing several features, including stemming and N-gram techniques. These features have significantly improved the accuracy of the proposed model in spam review identification. Moreover, the proposed model utilizes and compares the accuracy of four different classifiers, including Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) to further improve the accurate prediction of spam review.

III. REVIEW DATASET

A major challenge in building a spam review detection model through supervised learning is the collection of a labelled review dataset. Most of the existing spam review detection methods, using supervised learning, are based on pseudo-fake reviews rather than collecting spam reviews filtered from

[3] www.yelp.com/dataset
[4] www.dianping.com

different commercial review websites. Pseudo-fake reviews are either generated through manual annotation or by the Amazon Mechanical Turk (AMT). The AMT is a crowd sourcing marketplace for freelancers where individuals and companies can easily hire people to write reviews according to their requirements [29]. Manual annotation of spam reviews is a very difficult and challenging task [30]. Since spam reviews are difficult to identify, turkers have the same psychological state of mind as an actual fake reviewer. Spam review detection models constructed by using pseudo-fake reviews produce better accuracy in the training phase as compared to models constructed with the help of real-world spam reviews. However, such models that are trained on pseudo-fake reviews are not effective in detecting real-world spam reviews [4].

Considering the above issues, this work uses a real-world Amazon product review dataset[5] which includes the rich behavior and posting history of the reviewers. The dataset contains 26.7 million reviews, 15.4 million reviewers and 3.1 million products that primarily fell into six categories. Table 1 presents a detailed distribution of the dataset in terms of categories, reviews, reviewers and products. Spam review detection using the linguistic method requires a labelled dataset to train the classifier, but the Amazon product review dataset used in this study was not labelled. Therefore, to tackle this problem, the researchers first utilizes the SRD-BM (Section IV) to create a labelled dataset and then, the SRD-LM uses this labelled dataset to train the classifier. In SRD-LM, data-pre-processing, tokenization, review content analysis, feature extraction and selection, and classification is performed by using the Natural Language Toolkit, NLTK[6] 3.0, which provides easy to use built-in text processing libraries.

TABLE 1. Detailed distribution of Amazon dataset used in proposed method.

Category	Total Reviews	Total Reviewers	Total Products
Cell Phones and Accessories	3446396	2260636	319652
Clothing, Shoes, and Jewellery	5748260	3116944	1135948
Electronics	7820765	4200520	475910
Home and Kitchen	4252723	2511106	410221
Sports and Outdoor	3267538	1989985	478846
Toys and Games	2251775	1342419	327653
Total	26787457	15421610	3148230

IV. SPAM REVIEW DETECTION USING THE SPAMMER BEHAVIORAL METHOD (SRD-BM)

This section elaborates the proposed spammer behavioral method and analyzes the performance of the method in terms of accurate identification of spam reviews. Since a spammer can be identified by analyzing its different behavioral features, therefore, unlabelled dataset can be used with unsupervised learning to identify the spam reviews [31], [32]. The proposed Spammer Behavioral Method (SRD-BM) takes unlabelled dataset and produces an output of a labelled dataset that identifies spam and not spam reviews. This labelled dataset will be used as input in Section V for the proposed Spam Review Detection using Linguistic Method (SRD-LM).

A. PROPOSED FRAMEWORK OF SRD-BM

A framework of the proposed SRD-BM is shown in Figure 1. The process starts with the identification and calculation of spammer behavioral features in unlabelled Amazon review dataset. This calculation is performed on all reviews of the dataset based on the equations listed in Section IV-A.1. The average score of respective reviews in complete dataset is calculated using normalized values of each behavioral feature. This average score is then used to calculate accuracy of spam review identification using mean value method. Next, to identify the importance of each behavioral feature, the process continues by dropping each feature one-by-one and recalculates the updated average score, named as drop score. The accuracy achieved using average score is compared with that of drop score. If the achieved accuracy is dropped by 5% than a weight of "2" is assigned to that specific dropped behavioral feature otherwise a weight of "1" is assigned. Similarly, all behavioral features are assigned weights based on their importance in the dataset. Next, spam score of each review is calculated with respect to the assigned weights to each behavioral feature. This spam score is then compared with a variable threshold to highlight the review as spam or not spam. The complete process has been elaborated in the following sub-sections.

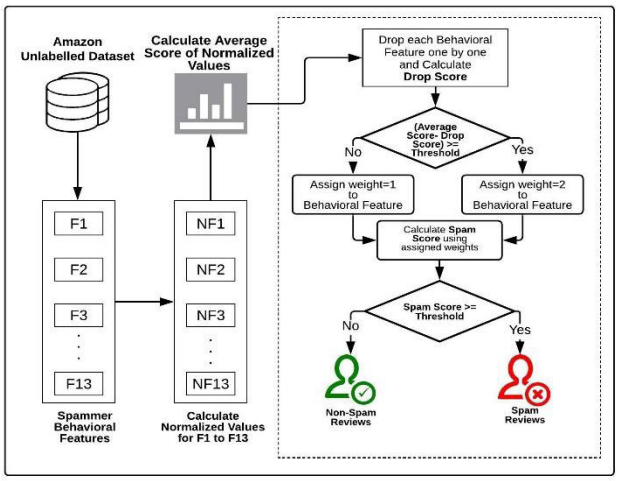

FIGURE 1. The framework of the spammer behavioral method (SRD-BM).

The proposed SRD-BM executes in four phases: (1) First it calculates the normalized value (0-1) of each spammer behavioral feature. (2) Based on these values, it calculates

[5]http://jmcauley.ucsd.edu/data/amazon/links.html
[6]www.nltk.org

the mean score for each review and the overall accuracy of the complete dataset. (3) Next, it assesses the impact of each behavioral feature by following dropping feature method and assigns a weight according to the importance of each behavioral feature. (4) Finally, it calculates spam score using weighted behavioral features and identifies spam and not-spam reviews using different threshold values.

1) SPAMMER BEHAVIORAL FEATURES

The reviewer's behavioral features point out the characteristics which are likely to be linked with a spammer and, thus, can be exploited to identify spam and non-spam reviews [33]. These features may act as indication to identify the spammers and must not be completely treated as a condition to mark a reviewer as spammer or not spammer. Therefore, the proposed approach uses a rich set of behavioral features and does not rely on a single behavioral feature for spammer identification. Eqs. (1), (2) .... (15) of each spammer behavioral feature are discussed in this section. Based on these equations, the normalized (0-1) values of each behavioral feature are calculated. The values close to 0 indicate not-spam whereas values nearer to 1 represent spam reviews. The notations used in this section are listed in Table 2.

TABLE 2. List of notations.

Variables	Description
a	Author 'a'
r	Review 'r'
r_a	Review 'r' by the author 'a'
R_a	Set of all reviews of an author 'a'
t	Time of current review 'r'
L_a, F_a	Time of the last review of the author 'a', Time to the first review of the author 'a'
*(r)	Rating of review 'r'
$len(r)$	Number of characters in a review 'r'
$MaxRev(a)$	Maximum number of reviews by the author 'a'

a: CONTENT SIMILARITY (CS) - F1

For their ease, spammers usually copy their reviews across similar products as they do not want to put much effort into creatively writing spam reviews [34]. Therefore, capturing the content similarity of the reviews of an author is important to detect their spamming behavior. This work used the cosine similarity to capture the content similarity of the reviews.

$$F_1(a) = \max_{x}^{y} [cosine(r_i, r_x)] \quad where\ r_i, r_x \in R_a,\ x < y \quad (1)$$

Here r_i is the current review and y is the total number of previous reviews by that author. The proposed method calculates the cosine similarity of each review with the previous review and selects the maximum value out of it.

b: MAXIMUM NUMBER OF REVIEWS (MNR) - F2

If any author posts too many reviews in a single day, then it may indicate an abnormal behavior. Hence, the proposed approach calculates the ratio of the total number of reviews of an author by the maximum number of reviews posted by that author in previous days.

$$F_2(a) = \frac{MaxRev(a)}{max(MaxRev(a))} \quad (2)$$

c: REVIEW BURSTINESS (RB) - F3

Most spammers tend to burst reviews to get fast results. Posting too many reviews in a short time is considered as unusual activity and may indicate a spammer. The proposed approach calculates the number of reviews by an author in the previous 24 hours. If the total count exceeded a threshold then the current review is more likely to be spam. Through the experimental analysis of the dataset, the threshold value is set to X = 12.

$$F_3(a) = \begin{cases} 1, & \sum_{x=1}^{|R_a|} |\{r_x \in R_a\} \cap (t_x\ is\ in\ last\ 24\ hours| > X) \\ 0, & otherwise \end{cases} \quad (3)$$

d: ACTIVITY WINDOW (AW) - F4

Spammers are usually not long-time members of any website. On the other hand, genuine reviewers post reviews from time to time. Thus, it is important to identify such authors. The proposed approach calculates the difference between the timestamps of the first and last reviews of an author to find out the number of active days an author had been on the system. Through the experimental analysis of the dataset, the threshold value is set to X = [0, 45].

$$F_4(a) = \begin{cases} 1, & L_a - F_a < X \\ 0, & otherwise \end{cases} \quad (4)$$

e: REVIEW COUNT (RC) - F5

Since spammers are not long-time members, as established through the activity window analysis, it was deduced that they were more likely to have a lower number of reviews than that of genuine authors [35]. Thus, the proposed approach analyzes the dataset to find out the number of reviews written by the authors. If a specific author writes less reviews than the threshold then it is considered as spammer. The threshold of X = 5, obtained through experimental analysis, is used to distinguish between fake and genuine authors.

$$F_5(a) = \begin{cases} 1, & |R_a| < X \\ 0, & otherwise \end{cases} \quad (5)$$

f: THE RATIO OF POSITIVE REVIEWS (PR) - F6

Spam reviews can be used for both promoting and demoting businesses. The proposed approach calculates the percentage of positive reviews (reviews with 4 and 5 ratings) by an author to filter out those authors who were more inclined towards promoting businesses. This ratio highlights that this respective author is inclined towards writing positive reviews

for a specific product.

$$F_6(a) = \frac{\sum_{x=1}^{|R_a|} |\{\star(r_x) \in \{4, 5\}\}|}{|R_a|} \quad (6)$$

g: THE RATIO OF NEGATIVE REVIEWS (NR) - F7

As the percentage of positive reviews is important, it is also important to find the percentage of negative reviews by any author. To find out the percentage of such reviews (reviews with 1 and 2 ratings), the proposed approach filtered out those authors who are more inclined towards demoting businesses by calculating the percentage of its negative reviews.

$$F_7(a) = \frac{\sum_{x=1}^{|R_a|} \left|\{\star(r_x) \in \{1, 2\}\}\right|}{|R_a|} \quad (7)$$

h: THE RATIO OF FIRST REVIEWS (FR) - F8

Early reviews on products or services can have a major impact on businesses. Thus, spammers try to become early reviewers as this enables them to be more influential [6]. The proposed approach calculates the ratio of the first review of an author by the total number of reviews by that author.

$$F_8(a) = \frac{\sum_{x=1}^{|R_a|} |\{r_x \in R_a\} \cap (r_x is\ a\ first\ review)|}{|R_a|} \quad (8)$$

i: REVIEW OF A SINGLE PRODUCT (RSP) - F9

Usually, the spammers write reviews for a single product for which they have been hired. They write reviews for a single product as their main objective could be to promote or demote that specific product's market. Using this feature, the proposed approach detects all such authors' who usually write reviews for a specific/single product and marked them as spam.

$$F_9(a) = \begin{cases} 1, & R_a \in (Single\ Product) \\ 0, & otherwise \end{cases} \quad (9)$$

j: RATING DEVIATION (RD) - F10

A genuine reviewer is anticipated to give a rating that is similar to the rating of any other reviewer of the same product. Mostly, a spammer gives a different rating from the genuine reviewer's ratings because the purpose of the spammer is a false projection of a product either in a positive or negative sense. Using this feature, this study calculates the mean rating value of the product by using Eq. (10). Moreover, based on the calculated mean value, the normalized spam score of the rating deviation feature is calculated by Eq. (11)

$$MEAN_r = \frac{\sum_{x=1}^{|R_a|} |\star(r_x)|}{|R_a|} \quad (10)$$

$$F_{10}(r) = \frac{|\star(r_a) - MEAN_r|}{4} \quad (11)$$

k: REVIEW LENGTH (RL) - F11

As spammers attempt to write fake experiences, therefore they do not have much content to write or it can be said that spammers usually do not invest much time in writing a single review. Analysis of the review dataset exhibits that on average reviewers write X = 400 characters. Using this value as a threshold, the proposed approach filters out such reviews as spam which have less than X characters. The reviewers of such reviews are thus marked as spammers.

$$F_{11}(a) = \begin{cases} 1, & len(r_a) < X \\ 0, & otherwise \end{cases} \quad (12)$$

l: EXTREME RATING (ER) - F12

Considering the rating attribute in the review dataset, the rating deviation from the mean rating is one parameter to identify the spammer. In the similar context, the proposed approach analyses another spammer behavioral feature, known as the extreme rating, to identify the spammer. It has been observed that spammers usually give an extreme rating (i.e., 5 stars or 1 star) as their main objective is to promote/demote products/businesses [36]. Eq. (13) filters out all such reviewers as spammers who have throughout given 1 or 5 stars rating.

$$F_{12}(a) = \begin{cases} 1, & \star(r_a) \in \{1, 5\} \\ 0, & \star(r_a) \in \{2, 3, 4\} \end{cases} \quad (13)$$

m: THE RATIO OF CAPITAL LETTERS (RCL) - F13

Spammers usually try to emphasize or get attention by using capital letters. Sentences with more ratios of capital letters may grab the attention of readers and it is also unusual to have many capitals in a sentence. The proposed approach counts the number of capital letters a specific reviewer has used Eq. (14) and filters them out as a spammer Eq. (15).

$$ECL(r_a) = |count\_caps(r_a) - count\_sent(r_a)| \quad (14)$$

$$F_{13}(a) = \begin{cases} 0, & ECL(r_a) = 0 \\ \dfrac{ECL(r_a)}{count_{l(r_a)}}, & otherwise \end{cases} \quad (15)$$

2) MEAN VALUE OF ALL NORMALIZED BEHAVIORAL FEATURES

Based on the values of the above behavioral features, the mean value is calculated using the normalized values of these features. The purpose is to assess the overall accuracy of spam review identification using mean values of all behavioral features. The accuracy achieved using mean value is then compared with that of drop score accuracy calculated with drop feature method.

3) DROPPING INDIVIDUAL SPAMMER BEHAVIORAL FEATURE METHOD

This method drops each behavioral feature, one by one, and compares the accuracy result obtained from mean score of all behavioral features with that of mean score with dropped feature. If the accuracy, obtained from mean score with the dropped feature, drops by 5% or more, then it assigns a weight of "2" to that feature, otherwise assigns a weight of "1". Hence, dropping feature method assesses the importance of each behavioral feature on the review dataset.

4) SPAM SCORE METHOD

This method calculates spam score in equation 16 using normalized value of each behavioral features and weights assigned to each behavioral feature through dropping feature method (16), as shown at the bottom of this page, where *a1, a2 ... a13* are the weights assigned through dropping feature method and *F1, F2 ... F13* represents the calculated normalized values of each behavioral feature. Eq. (16) calculates the spam score of each review in the dataset by dividing the sum of the multiplication of weights assigned to each behavioral feature with their normalized values with total weight. Next, the proposed method categorizes the reviews as spam or not spam based on the comparison of the respective spam score with a variable threshold. Through experimental evaluations, the variable threshold is set to $\tau = 0.5$, 0.55 and 0.6. Eq. (17) represents the process of assigning the label to each review from $\{L_{not-spam}, L_{Spam}\}$ where i represents the review number.

$$L_{r[i]} = \begin{cases} L_{normal} & Score_{r[i]} < \tau \\ L_{spam} & Score_{r[i]} >= \tau \end{cases} \quad (17)$$

The complete procedure of SRD-BM is implemented using the algorithm represented in Figure 2. Line 2-9 calculates the normalized values of each behavioral feature of all reviews and uses these to calculate mean value of all 13 behavioral features. Line 10-20 implements drop feature method and assigned weights to each behavioral feature based on their importance. Line 21-30 calculates spam score using assigned weights to each behavioral feature and identifies spam reviews of dataset using variable threshold.

B. RESULTS AND DISCUSSION

The performance of the proposed SRD-BM is analyzed using evaluation parameters of precision, recall, f-measure and accuracy. The Logistic Regression classifier is used for training and testing the labelled dataset obtained from proposed SRD-BM. This classifier is the best performing algorithm as analyzed using the proposed SRD-LM (Section V). Next, K-fold cross-validation (k = 5) is used to measure and validate the accuracy of the proposed SRD-BM. In k-fold cross-validation, the dataset is divided into k equal-sized segments. Through each run, one segment is considered for testing the proposed model and other k-1 segments are used for its training. This process is repeated k times so that each segment is used exactly once to train the proposed model. The accuracy is computed by considering the average accuracy of all runs. The experimental evaluation is performed in three phases. First, the accuracy is calculated using mean value of each review exploiting all spammer behavioral features. Next, to assess the impact of individual spammer behavioral features, the review dataset is analyzed by adapting drop feature method. Finally, overall accuracy using spam score method is calculated to assess the effectiveness of proposed SRD-BM in identifying the review as spam and not spam.

Algorithm 1: Spam review detection using behavioral features method

Input: review R_i , $\tau = 0.5,0.55,0.6$ //threshold value for labelling the review
Output: Spam or Not-Spam

```
1.  for each review Ri in review dataset do
2.      // behavior features (F1, F2, F3, ..., F13)
3.      for each behavior feature Fi calculate normalize value do
4.          // variable V is calculating normalize value of F
5.          Vi = calculate normalize value Fi
6.          Sum += Vi
7.      end for
8.      // calculating average score
9.          Average Score = Sum / 13
10.     for each value Vi do
11.         // calculating drop score
12.         DropScore = (Sum - Vi) / 12
13.         if | Average Score – DropScore | >= 0.05 then
14.             assign weight Wi ◄— 2
15.             Total Weight += 2
16.         else
17.             assign weight Wi ◄— 1
18.             Total Weight += 1
19.         end if
20.     end for
21.     for each value Vi do
22.         // calculating total spam score
23.         Score += Wi * Vi
24.     end for
25.     Spam Score = Score / Total Weight
26.     if Spam Score > τ then
27.         label Ri ◄— Spam
28.     else
29.         label Ri ◄— Not-Spam
30.     end if
31. end for
```

FIGURE 2. Algorithm of proposed SRD-BM.

1) PERFORMANCE ANALYSIS OF MEAN VALUE METHOD

Table 3 represents the experimental results of the mean value method using all behavioral features. All evaluation parameters are calculated using different threshold values i.e. 0.5, 0.55 and 0.6, to identify each spam review. It has been observed that the highest accuracy of 0.861 is achieved using threshold value $\tau = 0.50$ when compared with that of other threshold values. This shows that threshold value of 0.5 gives balanced threshold for spam review identification.

TABLE 3. Performance analysis of mean value method.

Feature Setting	Threshold (τ)	Precision	Recall	F-measure	Accuracy
Behavioral Features	0.50	0.89	0.86	0.87	**0.861**
	0.55	0.87	0.85	0.86	0.842
	0.60	0.84	0.82	0.83	0.81

2) PERFORMANCE ANALYSIS OF DROPPING INDIVIDUAL SPAMMER BEHAVIORAL FEATURE METHOD

To investigate the contribution of each spammer behavioral feature, the analysis of SRD-BM is performed using a drop feature method. Table 4 shows the accuracy results

$$Spam\ Score = \frac{(a1F1 + a2F2 + a3F3 + a4F4 + a5F5 + a6F6 + a7F7 + a8F8 + a9F9 + a10F10 + a11F11 + a12F12 + a13F13)}{\sum_{k=1}^{13} ak} \quad (16)$$

TABLE 4. Performance analysis of drop feature method.

Dropped Features	Precision	Recall	F-measure	Accuracy
CS	0.843	0.804	0.810	**0.804**
MNR	0.813	0.884	0.807	**0.793**
RB	0.886	0.865	0.866	0.854
AW	0.803	0.757	0.774	**0.758**
RC	0.812	0.806	0.801	**0.799**
PR	0.794	0.776	0.776	**0.765**
NR	0.866	0.858	0.857	0.849
FR	0.856	0.836	0.867	0.832
RSP	0.830	0.808	0.828	**0.808**
RD	0.847	0.810	0.826	0.816
RL	0.842	0.807	0.816	**0.807**
ER	0.864	0.830	0.842	0.836
RCL	0.860	0.820	0.840	0.828

using review dataset by dropping each behavioral features one-by-one. The accuracy calculated by dropping each feature individually is compared with mean value accuracy (Section IV-B.1) which assesses the impact of dropping that specific feature. It has been observed that the accuracy results of dropping the CS, MNR, RC, PR, RSP, AW and RL features are reduced by 5% or more when compared with the mean value method. On the other hand, dropping other behavioral features RB, NR, FR, RD, ER, and RCL reduces the accuracy by only 2-4%. This shows that the CS, MNR, RC, PR, RSP, AW and RL are comparatively significant spammer behavioral features. Therefore, the drop feature method assigns a weight of "2" to CS, MNR, RC, PR, RSP, AW and RL and weight of "1" to RB, NR, FR, RD, ER, and RCL spammer behavioral features.

3) PERFORMANCE ANALYSIS OF SRD-BM USING SPAM SCORE

The proposed SRD-BM calculates the spam score of each review by using Eq. 16 and 17. Table 5 shows the accuracy results of spam review identification in the review dataset by applying proposed SRD-BM. It has been observed that proposed model produces better accuracy with the threshold value of 0.5 when compared with that of other threshold values. It can also be deduced that the threshold value (τ) can be variable depending upon different applications. For example, when an application wants to identify as many spam reviews as possible, then he or she ought to set τ to be relatively small.

TABLE 5. Performance analysis of SRD-BM using spam score.

Feature Setting	Threshold (τ)	Precision	Recall	F-measure	Accuracy
Behavioral Features	0.5	0.95	0.93	0.94	**0.931**
	0.55	0.90	0.90	0.88	0.892
	0.60	0.88	0.87	0.86	0.876

C. PERFORMANCE ANALYSIS OF SRD-BM UTILIZING DIFFERENT DATASET

In this section, the performance of SRD-BM is analysed by utilizing Yelp dataset.[7] The dataset contains 1,035,045 review, 458,565 reviewers of 6,168 hotels and restaurants. Table 6 represent the comparison of SRD-BM with Kumar *et al.* [24] and Zhang *et al.* [25]. These existing approaches utilized Yelp dataset and analysed different spammer behavioural features. Table shows that SRD-BM outperforms the existing approaches by achieving an accuracy of 92% on Yelp dataset. This improvement in accuracy is the result of incorporating rich set of spammer behavioural features while calculating spam score of each review.

[7]http://odds.cs.stonybrook.edu/yelpzip-dataset

TABLE 6. Comparative Analysis of SRD-BM with existing approaches using Yelp dataset.

Existing Studies	Dataset	Accuracy	Proposed SRD-BM Accuracy
Kumar et al. [24]	Yelp	81%	**92%**
Zhang et al [25]		87%	

V. SPAM REVIEW DETECTION USING LINGUISTIC METHOD (SRD-LM)

It has been observed from the literature that Spam Review detection using linguistic method uses only review text for spotting the spam review [37], [38]. It is usually performed binary classification in which the review is classified as "spam" or "not spam". This section elaborates the proposed SRD-LM. It describes the process of feature extraction and selection from the review text. It also describes different classification algorithms that are used to train and test the proposed method.

A. PROPOSED SRD-LM

The proposed linguistic method uses different data preprocessing techniques, transformation, feature selection and machine learning classification algorithms to develop an accurate spam review detection model. The complete process for the proposed SRD-LM executes in six steps which have been presented in Figure 3.

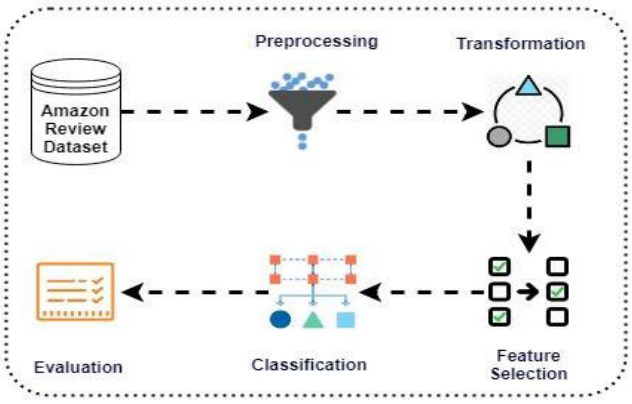

FIGURE 3. Process of SRD-LM.

a) **Dataset**: The first step of the proposed SRD-LM is to input labelled Amazon review dataset. This labelled

TABLE 7. Evaluation of SRD-LM with Naive Bayes classifier.

Method		Precision	Recall	F-Measure	Accuracy	AUROC
Unigram	IG (top 1%)	0.866	0.847	0.856	84.036	0.853
	IG (top 2%)	0.843	0.821	0.847	82.282	0.843
	IG (top 3%)	0.828	0.802	0.818	80.170	0.825
	IG (top 4%)	0.791	0.771	0.789	77.037	0.806
Bigram	IG (top 1%)	0.871	0.853	0.867	**85.864**	0.888
	IG (top 2%)	0.874	0.853	0.853	85.154	0.871
	IG (top 3%)	0.840	0.824	0.830	82.480	0.855
	IG (top 4%)	0.839	0.811	0.827	81.040	0.830
Trigram	IG (top 1%)	0.856	0.838	0.827	83.507	0.858
	IG (top 2%)	0.822	0.805	0.818	80.080	0.831
	IG (top 3%)	0.812	0.794	0.808	79.983	0.820
	IG (top 4%)	0.796	0.775	0.790	77.047	0.780
Unigram + Bigram	IG (top 1%)	0.830	0.810	0.824	81.067	0.841
	IG (top 2%)	0.817	0.780	0.799	78.321	0.806
	IG (top 3%)	0.811	0.793	0.805	79.334	0.801
	IG (top 4%)	0.807	0.781	0.775	78.154	0.797
Bigram + Trigram	IG (top 1%)	0.740	0.729	0.730	72.910	0.749
	IG (top 2%)	0.732	0.718	0.725	71.324	0.733
	IG (top 3%)	0.700	0.690	0.690	69.510	0.715
	IG (top 4%)	0.673	0.669	0.670	66.100	0.682
Unigram + Bigram + Trigram	IG (top 1%)	0.663	0.652	0.668	65.554	0.676
	IG (top 2%)	0.654	0.641	0.658	64.832	0.668
	IG (top 3%)	0.656	0.641	0.658	64.326	0.654
	IG (top 4%)	0.630	0.610	0.620	61.118	0.649

dataset is obtained after applying proposed SRD-BM (Section IV) This dataset will be considered for the training and testing of the proposed SRD-LM model.

b) **Pre-processing:** The pre-processing of the dataset is performed using the following methods:

- **Removing Stop Words or Punctuation:** Generally, the review text contains unnecessary words like ''is'', ''the'', ''and'', ''a''. These words are not helpful in detecting spam reviews; so, it is better to remove them before tokenization to avoid noise and unnecessary tokens. For example, consider a review ''This is a very good car'' after removing stop words and punctuation, the review looks like ''good car''.
- **Stemming Word:** A stemming algorithm converts different forms of the word into a single recognizable form. For example, consider the words ''works'', ''working'', ''worked'' are used as an instance of the word ''work''. Stemming is applied to the review text before tokenizing it to make the review more compact and understandable.
- **Tokenizing:** Tokenizing is an important pre-processing step [26] and is used for splitting text into individual words or sequences of words. For example, consider a review ''good car''. When we apply the N-gram tokenizing technique on it, we have several different combinations. i.e. Uni-gram: [''good'', ''car''], Bi-gram: [''good car''], Uni + Bi-gram: [''car'', ''good'', ''good car'']. In a similar way, the tri-gram technique uses three words as token. The proposed SRD-LM employs different N-gram combinations on the review data.

c) **Transformation:** The linguistic method works on numeric data. Therefore, to convert textual review data to the numeric form, the proposed SRD-LM uses a document term matrix [39]. A document term matrix is used to represent tokens generated by an N-gram model in the form of a sparse matrix. The sparse matrix defines the frequency of the terms or tokens in the collection of reviews. In this work, the TF-IDF has been applied to transform the review text into the numerical vector. Moreover, the TF-IDF reflects the importance of each word or term in the collection of reviews. The TF-IDF value increases according to the number of times a token appears in a document. The TF-IDF is one of the most popular term-weighting schemes and provides better results than a simple count technique.

d) **Feature Selection:** Feature selection technique is used to select most important features which appear in the review dataset. The proposed approach uses Information Gain (IG) for feature selection. Term frequency has been used to select the top 1%, 2%, 3% and 4% features, respectively.

TABLE 8. Evaluation of SRD-LM with logistic regression classifier.

Method		Precision	Recall	F-Measure	Accuracy	AUROC
Unigram	IG (top 1%)	0.842	0.845	0.853	84.289	85.815
	IG (top 2%)	0.832	0.820	0.823	82.289	84.791
	IG (top 3%)	0.823	0.819	0.821	81.113	82.791
	IG (top 4%)	0.813	0.809	0.812	80.006	81.784
Bigram	IG (top 1%)	0.864	0.850	0.860	85.102	86.373
	IG (top 2%)	0.847	0.830	0.842	83.595	84.943
	IG (top 3%)	0.825	0.834	0.833	83.425	84.593
	IG (top 4%)	0.813	0.806	0.818	82.395	83.341
Trigram	IG (top 1%)	0.807	0.785	0.791	78.988	80.860
	IG (top 2%)	0.799	0.786	0.790	78.745	79.952
	IG (top 3%)	0.791	0.779	0.783	77.705	78.565
	IG (top 4%)	0.785	0.779	0.772	77.554	78.647
Unigram + Bigram	IG (top 1%)	0.881	0.873	0.887	**88.523**	89.583
	IG (top 2%)	0.880	0.871	0.885	87.346	88.304
	IG (top 3%)	0.877	0.862	0.875	86.293	87.156
	IG (top 4%)	0.858	0.844	0.846	84.156	86.038
Bigram + Trigram	IG (top 1%)	0.752	0.757	0.761	75.338	73.473
	IG (top 2%)	0.748	0.741	0.758	74.585	73.976
	IG (top 3%)	0.759	0.748	0.754	74.412	74.603
	IG (top 4%)	0.754	0.740	0.751	74.365	73.308
Unigram + Bigram + Trigram	IG (top 1%)	0.808	0.795	0.798	79.476	80.651
	IG (top 2%)	0.800	0.788	0.792	79.313	80.356
	IG (top 3%)	0.807	0.799	0.803	79.283	80.179
	IG (top 4%)	0.801	0.791	0.794	79.183	80.040

e) **Classification Methods**: After the text reviews are converted to the document-term matrix, these matrixes are considered as input for the following four supervised learning algorithms for the classification purpose.

- **Naïve Bayes (NB) Classifier:** Naïve Bayes (NB) classifier, also called a linear classifier, is used for both classifications as well as training purposes. This is a probabilistic classifier method based on Bayes' theorem. Naïve Bayes classifiers are based upon the naïve assumption that the features in a dataset are mutually independent. The following equation is the mathematical representation of the Naïve Bayes classifier.

$$P(C \mid X) = \frac{P(X \mid C)\, P(C)}{P(x)}$$

$P(C \mid X)$ is the posterior probability of the target class with a given predicate attribute. $P(C)$ is the prior probability of class. $P(X \mid C)$ is the probability of the predictor class. $P(x)$ is the prior probability of the predictor. This study uses Bernoulli Naïve Bayes and the feature vector is represented by 0 and 1, where 0 indicates a feature that does not occur in the review and 1 represents a feature that occurs in the review.

- **Logistic Regression (LR) classifier:** Logistic Regression (LR) is a statistical method for analyzing a dataset. This classifier uses one or more independent variables to determine the result. The results are measured with a binary variable either 0 or 1. LR assumes that the posterior distribution, $P(y|x)$, takes the shape of a logistic function. Here y is the label and x is the set of features.
- **Support Vector Machine (SVM) classifier:** The Support Vector Machine (SVM) is a state of art classifier and its optimization procedure maximizes the predictive accuracy while automatically avoiding over-fitting of the training data. The SVM projects the input data into kernel space then builds a linear model in this kernel space. The SVM is the standard tool for machine learning and data mining.
- **Random Forest (RF) classifier**: A Random Forest (RF) is a meta estimator that fits many decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It seems to be quite popular these days due to its several advantages, such as being faster and more scalable compared to other machine learning models.

f) **Evaluation**: Finally, the proposed SRD-LM is implemented and evaluated using a different variation of the N-gram model and above-mentioned classification algorithms. The complete steps are also depicted in Figure 3.

TABLE 9. Evaluation of SRD-LM with support vector machine classifier.

Method		Precision	Recall	F-Measure	Accuracy	AUROC
Unigram	IG (top 1%)	0.880	0.850	0.876	**86.525**	88.428
	IG (top 2%)	0.861	0.858	0.864	85.352	86.433
	IG (top 3%)	0.862	0.844	0.852	84.482	86.546
	IG (top 4%)	0.855	0.842	0.851	83.352	85.638
Bigram	IG (top 1%)	0.852	0.837	0.841	83.518	86.571
	IG (top 2%)	0.849	0.820	0.838	82.240	84.916
	IG (top 3%)	0.836	0.818	0.824	81.126	83.210
	IG (top 4%)	0.825	0.800	0.818	80.016	82.008
Trigram	IG (top 1%)	0.855	0.837	0.846	83.038	85.084
	IG (top 2%)	0.838	0.818	0.829	81.186	83.974
	IG (top 3%)	0.827	0.806	0.817	80.010	82.816
	IG (top 4%)	0.816	0.799	0.816	79.390	81.436
Unigram + Bigram	IG (top 1%)	0.840	0.821	0.830	82.589	84.350
	IG (top 2%)	0.803	0.833	0.824	81.978	83.292
	IG (top 3%)	0.807	0.837	0.832	81.301	83.598
	IG (top 4%)	0.795	0.839	0.828	81.171	72.259
Bigram + Trigram	IG (top 1%)	0.847	0.820	0.839	82.072	84.791
	IG (top 2%)	0.843	0.812	0.822	81.150	83.089
	IG (top 3%)	0.823	0.808	0.816	80.073	82.268
	IG (top 4%)	0.772	0.762	0.782	76.846	80.943
Unigram + Bigram + Trigram	IG (top 1%)	0.831	0.810	0.825	81.196	83.637
	IG (top 2%)	0.800	0.780	0.792	78.345	79.563
	IG (top 3%)	0.759	0.750	0.762	75.294	78.706
	IG (top 4%)	0.745	0.729	0.731	72.151	73.245

B. RESULTS AND DISCUSSION

The proposed SRD-LM is evaluated from the following two perspectives: (1) Evaluation of SRD-LM using different combinations of N-gram features, the variation of Information Gain (IG) for feature selection and four classification algorithms (NB, LR, SVM, RF) in terms of accuracy in spam review detection. (2) Comparison of SRD-LM with existing linguistic techniques of spam review identification. These evaluation results have been presented in the following sub sections.

1) EVALUATION OF CLASSIFICATION ALGORITHMS USING SRD-LM

In this section, four classification algorithms are evaluated using SRD-LM with different N-gram combinations coupled with various IG variations. These classification algorithms are compared in terms of achieved accuracy over review dataset, and the one giving better results is considered as the most accurate algorithm in spam review identification.

a: SRD-LM WITH NAÏVE BAYES CLASSIFICATION

SRD-LM is evaluated with different combinations of the N-gram features and IG using Naïve Bayes classification in terms of precision, recall, f-measure, accuracy and AUROC parameters. The impact of these different combinations is shown in Table 7. It is observed from the experimental results that the maximum accuracy of 85.864 is achieved when the Naïve Bayes classifier is implemented with a combination of bigram with IG (top 1%). It can also be observed that the accuracy value obtained using bi-gram is better than that of uni-gram and tri-gram. The reason being Naïve Bayes classifier is based on a probabilistic technique, where the features are independent of each other. Hence, when the analysis is carried out using uni-gram and bi-gram, the accuracy value is better than that of tri-gram. In tri-gram approach, the words are repeated several times; thus, it affects the probability of the document. For example, consider the review about a product ''It is not a bad product'', after using the tri-gram technique ''It is not'' and ''is not a'' show negative popularity whereas the review about the product represents a positive sentiment. Therefore, the accuracy of the classification decreases. It can also be observed from the analysis that when tri-gram is combined with uni-gram and bi-gram, it makes the accuracy values comparatively low.

b: SRD-LM WITH LOGISTIC REGRESSION (LR) CLASSIFICATION

SRD-LM is evaluated using the N-gram technique and different combinations of feature selection (IG) with Logistic Regression classifier. The evaluation is performed using different evaluation measures e.g. precision, recall, f-measure and AUROC. The results are presented in Table 8. It is

TABLE 10. Evaluation of SRD-LM with random forest classifier.

Method		Precision	Recall	F-Measure	Accuracy	AUROC
Unigram	IG (top 1%)	0.750	0.732	0.741	73.318	76.185
	IG (top 2%)	0.755	0.731	0.742	73.268	75.650
	IG (top 3%)	0.743	0.738	0.721	72.428	74.733
	IG (top 4%)	0.740	0.729	0.735	72.588	74.840
Bigram	IG (top 1%)	0.844	0.827	0.831	82.596	85.185
	IG (top 2%)	0.840	0.823	0.833	82.293	84.051
	IG (top 3%)	0.828	0.822	0.832	81.273	83.035
	IG (top 4%)	0.824	0.826	0.839	81.810	82.921
Trigram	IG (top 1%)	0.840	0.822	0.835	82.715	86.612
	IG (top 2%)	0.827	0.800	0.810	80.609	85.700
	IG (top 3%)	0.803	0.786	0.797	78.890	73.186
	IG (top 4%)	0.791	0.777	0.785	77.276	80.790
Unigram + Bigram	IG (top 1%)	0.857	0.832	0.842	83.068	86.097
	IG (top 2%)	0.832	0.815	0.820	81.528	82.245
	IG (top 3%)	0.829	0.808	0.801	80.452	82.138
	IG (top 4%)	0.803	0.781	0.799	78.692	79.295
Bigram + Trigram	IG (top 1%)	0.851	0.834	0.842	83.508	85.787
	IG (top 2%)	0.838	0.822	0.822	81.543	83.277
	IG (top 3%)	0.822	0.813	0.813	80.386	81.012
	IG (top 4%)	0.810	0.801	0.803	79.693	80.963
Unigram + Bigram + Trigram	IG (top 1%)	0.865	0.842	0.853	**84.037**	88.601
	IG (top 2%)	0.859	0.841	0.840	83.412	86.709
	IG (top 3%)	0.826	0.814	0.816	80.825	84.869
	IG (top 4%)	0.806	0.796	0.796	78.765	81.577

observed from the experimental results that the maximum accuracy of 88.523 is achieved when Logistic Regression classifier is implemented with a combination of unigram + bigram with IG (top 1%).

The Logistic Regression classifier is also a linear classifier as the decision boundary is determined by a linear function of the features. The proposed study uses binary classification (spam or not-spam) to determine the class label (using a threshold of 0.5). Through analysis of the results, it is observed that the Logistic Regression performed better if the variables are independent of each other. Therefore, uni-gram, bi-gram and a combination of uni-gram and bi-gram produced better results as compared to tri-gram and/or different combinations of tri-gram.

c: SRD-LM WITH SUPPORT VECTOR MACHINE (SVM) CLASSIFICATION

SRD-LM is evaluated with the N-gram technique and different combinations of IG using the support vector machine classifier in terms of precision, recall, f-measure, accuracy and AUROC. The analysis of the results is elaborated in Table 9. It is observed from the experimental results that the maximum accuracy of 86.525 is achieved when SVM classifier is implemented with a combination of unigram with IG (top 1%).

The SVM is a linear classifier and train the model to find a hyper plane to separate the reviews of dataset. As the uni-gram technique uses a single word, thus produces a better result using SVM. In the bi-gram and trigram techniques, different combinations of words are used. Therefore, when plotted in a hyper plane, they confuse the classifier and produce less accurate results as compared to the uni-gram technique. It can also be observed through the analysis that the combinations of uni-gram with bi-gram and tri-gram also produces less accuracy.

d: SRD-LM WITH RANDOM FOREST (RF) CLASSIFICATION

SRD-LM is evaluated with different combinations of N-gram and IG using the Random Forest classifier in terms of different evaluation measures of precision, recall, f-measure, accuracy and AUROC. The detailed results of experimental evaluations are presented in Table 10. It is observed from the experimental results that the maximum accuracy of 84.037 is achieved when the RF classifier is implemented with a combination of unigram + bigram + trigram with IG (top 1%). The Random Forest classifier builds a randomized decision tree and often produces good predictors. Moreover, each tree gives the classification and "votes" for that specific class. The forest chooses the classification having the majority vote (over all the trees in the forest). It can be observed from

TABLE 11. Comparative analysis of SRD-LM with existing linguistic approaches using Amazon reviews dataset.

Method		Rui Xia et al. [40]	Srikumar K [41]	Yan Dang et al. [42]	Rodrigo Moraes et al. [43]	G.Vinodhin et al. [44]	Proposed SRD-LM Method
Naïve Bayes	Unigram	80.9	⊗	81.3	76.2	79	**84.0**
	Bigram	82.8	71.27	⊗	⊗	⊗	**85.8**
	Trigram	⊗	⊗	⊗	⊗	⊗	83.5
	Unigram + Bigram	⊗	⊗	⊗	⊗	79.5	**81.0**
	Bigram + Trigram	⊗	⊗	⊗	⊗	⊗	72.9
	Unigram + Bigram +Trigram	⊗	⊗	⊗	⊗	⊗	65.5
Logistic Regression	Unigram	⊗	⊗	⊗	80.9	65.1	**84.2**
	Bigram	⊗	⊗	⊗	⊗	⊗	85.1
	Trigram	⊗	⊗	⊗	⊗	⊗	78.9
	Unigram + Bigram	⊗	⊗	⊗	⊗	65.3	**88.5**
	Bigram + Trigram	⊗	⊗	⊗	⊗	⊗	75.3
	Unigram + Bigram +Trigram	⊗	⊗	⊗	⊗	69.7	**79.4**
Support Vector Machine	Unigram	79.9	⊗	⊗	⊗	81	**86.5**
	Bigram	79.5	77.81	⊗	⊗	⊗	**83.5**
	Trigram	⊗	⊗	⊗	⊗	⊗	83.0
	Unigram + Bigram	⊗	⊗	⊗	⊗	80.3	**82.5**
	Bigram + Trigram	⊗	⊗	⊗	⊗	⊗	82.0
	Unigram + Bigram +Trigram	⊗	⊗	⊗	⊗	80.5	**81.1**
Random Forest	Unigram	⊗	⊗	⊗	⊗	⊗	73.3
	Bigram	⊗	81.3	⊗	⊗	⊗	**82.5**
	Trigram	⊗	⊗	⊗	⊗	⊗	82.7
	Unigram + Bigram	⊗	⊗	⊗	⊗	⊗	88.0
	Bigram + Trigram	⊗	⊗	⊗	⊗	⊗	83.5
	Unigram + Bigram +Trigram	⊗	⊗	⊗	⊗	⊗	84.0

the analysis that the Random Forest classifier was capable of easily handling the interactions between different features. Therefore, a combination of uni-gram, bi-gram and tri-gram produces better results as compared to individual uni-gram, bi-gram and tri-gram.

The analysis of the experimental evaluations of four classification algorithms using the Amazon review dataset is performed using SRD-LM for spam review detection. The process used different N-gram combinations with top features' selection using IG method. It is observed that the Logistic Regression algorithm performed best with the uni-gram + bi-gram combination in terms of accuracy and produced better ROC curve results as compared to the Naïve Bayes, Support Vector Machine and Random Forest classifiers.

2) PERFORMANCE EVALUATION OF SRD-LM WITH EXISTING LINGUISTIC METHODS

The comparative analysis based on the results obtained using the proposed SRD-LM to that of other linguistic techniques to identify spam review using the Amazon.com product review is presented in Table 11. The proposed SRD-LM is compared with five exiting linguistic methods. Xia *et al.* [40] used the Naïve Bayes and Support Vector Machine classifiers using the uni-gram and bi-gram approaches. Krishnamoorthy [41] used the Naïve Bayes, Support Vector Machine and Random Forest classifiers using the bi-gram approach. Dang *et al.* [42] used the Naïve Bayes classifier using the uni-gram approach. Moraes *et al.* [43] used the Naïve Bayes and Logistic Regression classifiers using the uni-gram approach to classify spam and not-spam reviews. Vinodhini and Chandrasekaran [44] implemented the Naïve Bayes, Logistic Regression and Support Vector classifiers with the uni-gram, a combination of uni-gram with bi-gram and a combination of uni-gram, bi-gram and tri-gram approaches. It can be observed from Table 11 that most of these existing approaches analyzed their models using only unigram and/or bigram techniques, whereas the proposed study analyzed SRD-LM using unigram, bigram, trigram and all possible combinations. It can also be observed that overall SRD-LM outperformed all the listed existing techniques.

C. DISCUSSION OF CLASSIFIERS IN VIEW OF SRD-LM

As per the experimental evaluation of SRD-LM, it is observed that the LR performed better than the other three classifiers i.e., SVM, NB and RF whereas SVM remained better than NB and RF, while NB achieves better accuracy than RF. The LR uses threshold values to determine the review as being spam or not-spam. SVM uses an absolute prediction

of 0 or 1. It is a linear classifier and trains the model to find a hyper plane to separate the reviews of dataset. NB is a probabilistic classifier; works on independent variables and performs better on a large dataset. RF forms several trees; therefore, it consumes more memory and had to slow down to make the evaluation. The RF classifier behaves like a black box that is very hard to understand and is considered very unpredictable in terms of accuracy. Therefore, based on these features and evaluation results LR produced overall better results as compared to all other classifiers using Amazon product review dataset.

VI. COMPARATIVE ANALYSIS OF SRD-BM AND SRD-LM

This section presents a comparative analysis of proposed SRD-BM and SRD-LM in terms of identifying spam reviews in dataset. Figure 4 demonstrates the results of this comparison in terms of identified spam reviews. It is evident that SRD-BM identified more spam reviews with better accuracy in dataset than SRD-LM. As mentioned in section III, total reviews of Amazon dataset were 26.7 million. It has been observed that SRD-BM efficiently identified 7,500,472 reviews as spam which formulates a proportion of 28% of the total reviews. Remaining 72% reviews are identified as not spam. On the other hand, SRD-LM, utilizing same Amazon dataset, identified 5,625,354 reviews as spam which makes the proportion of 21% of the total review dataset. This shows that SRD-BM is more accurate than SRD-LM in identifying spam reviews in large-scale real-world Amazon dataset. This has also been observed through experimental evaluation that SRD-BM achieved 93.1% accuracy whereas SRD-LM achieved 88.5% accuracy in spam review detection.

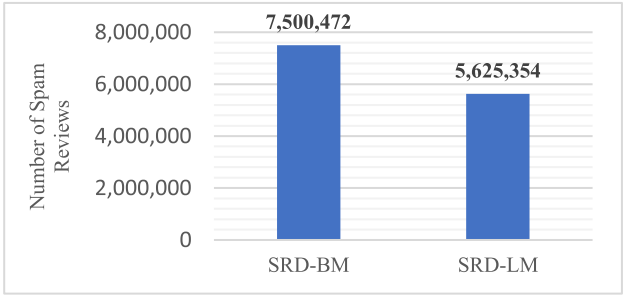

FIGURE 4. Comparison of SRD-BM and SRD-LM in terms of identified spam reviews.

VII. CONCLUSION

Online review spamming is a rapidly growing problem. Spam Review Detection (SRD) is a significant but challenging task as it is very difficult to differentiate the spam review from not-spam reviews. So far, many research works have attempted to identify the spammer and spam reviews, but these works have not been able to fully solve the spam review detection problem. This work performed an in-depth investigation of Amazon real-world dataset using the spammers' behavioral features and proposed SRD-BM and SRD-LM methods to detect spam reviews using behavioral and linguistic approaches respectively. To the best of the researcher's knowledge, this is the first study that analyzed and applied a rich set of spammers' behavioral features on a large-scale real-world review dataset. Furthermore, the experimental evaluation showed that the behavioral feature like content similarity, maximum number of reviews, review count, ratio of positive review, review of single product, activity window and review length features significantly improved the accuracy of the proposed SRD-BM. On the other hand, the proposed linguistic method SRD-LM, used N-gram techniques, transformation and feature selection, and different classification algorithms to further analyze the dataset for spam review detection. Through performance evaluation of each classifier, it is observed that the Logistic Regression performed better than the Support Vector Machine, Naïve Bayes and Random Forest. The comparison of the two proposed models indicated that the SRD-BM achieved better accuracy than the SRD-LM because SRD-BM uses behavioral attributes of dataset such as time stamps and ratings which provides additional support to identify spammers and thus spam reviews.

The findings of this study provide a practical implication for improving the trustworthiness of online product and service review platforms. The applications of the study include spam review detection in product/services reviews on e-commerce websites, product/services websites e.g. Amazon, Yelp, TripAdvisor, Daraz.pk, foodpanda.pk, etc. Future research will be focused on the availability of standard labelled datasets to train the classifiers. Furthermore, additional attributes will be added to the dataset to improve the accuracy and reliability of the spam review detection models. These may include an IP address of the spammer, registered an email address and signed-in location of the reviewer. Other future directions may be to identify spam reviews in multilingual review dataset and recognizing the spammer by feedback analysis of other users on their written reviews. A significant future direction of this work is to implement this problem utilizing deep-learning classifiers.

REFERENCES

[1] J. Huang, T. Qian, G. He, M. Zhong, and Q. Peng, "Detecting professional spam reviewers," in *Proc. Int. Conf. Adv. Data Mining Appl.* Berlin, Germany: Springer, 2013, pp. 288–299.
[2] S. Bajaj, N. Garg, and S. K. Singh, "A novel user-based spam review detection," *Procedia Comput. Sci.*, vol. 122, pp. 1009–1015, Jan. 2017.
[3] J. G. Biradar, S. P. Algur, and N. H. Ayachit "Exponential distribution model for review spam detection," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 3, pp. 938–947, 2017.
[4] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What yelp fake review filter might be doing?" in *Proc. Int. Conf. Web Social Media*, Jul. 2013, pp. 409–418.
[5] Y. Ren and D. Ji, "Learning to detect deceptive opinion spam: A survey," *IEEE Access*, vol. 7, pp. 42934–42945, 2019.
[6] T. Ong, M. Mannino, and D. Gregg, "Linguistic characteristics of shill reviews," *Electron. Commerce Res. Appl.*, vol. 13, no. 2, pp. 69–78, Mar. 2014.
[7] I. Dematis, E. Karapistoli, and A. Vakali, "Fake review detection via exploitation of spam indicators and reviewer behavior characteristics," in *Proc. Int. Conf. Current Trends Theory Pract. Inform.* Cham, Switzerland: Edizioni Della Normale, 2018, pp. 581–595.

[8] H. Li, Z. Chen, A. Mukherjee, B. Liu, and J. Shao, "Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns," in *Proc. Int. Conf. Web Social Media*, 2015, pp. 634–637.

[9] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*. Boston, MA, USA: Springer, 2012, pp. 415–463.

[10] S. Zhou, Z. Qiao, Q. Du, G. A. Wang, W. Fan, and X. Yan, "Measuring customer agility from online reviews using big data text analytics," *J. Manage. Inf. Syst.*, vol. 35, no. 2, pp. 510–539, Apr. 2018.

[11] L. Chen, L. Chun, L. Ziyu, and Z. Quan, "Hybrid pseudo-relevance feedback for microblog retrieval," *J. Inf. Sci.*, vol. 39, no. 6, pp. 773–788, Dec. 2013.

[12] C. Lin, Z. Huang, F. Yang, and Q. Zou, "Identify content quality in online social networks," *IET Commun.*, vol. 6, no. 12, pp. 1618–1624, 2012.

[13] M. Chakraborty, S. Pal, R. Pramanik, and C. R. Chowdary, "Recent developments in social spam detection and combating techniques: A survey," *Inf. Process. Manage.*, vol. 52, no. 6, pp. 1053–1073, Nov. 2016.

[14] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, "Spotting opinion spammers using behavioral footprints," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2013, pp. 632–640.

[15] A. Heydari, M. Tavakoli, and N. Salim, "Detection of fake opinions using time series," *Expert Syst. Appl.*, vol. 58, pp. 83–92, Oct. 2016.

[16] S. Kc and A. Mukherjee, "On the temporal dynamics of opinion spamming: Case studies on yelp," in *Proc. 25th Int. Conf. World Wide Web (WWW)*, 2016, pp. 369–379.

[17] H. Li, G. Fei, S. Wang, B. Liu, W. Shao, A. Mukherjee, and J. Shao, "Bimodal distribution and co-bursting in review spam detection," in *Proc. 26th Int. Conf. World Wide Web (WWW)*, 2017, pp. 1063–1072.

[18] N. Jindal and B. Liu, "Analyzing and detecting review spam," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, Oct. 2007, pp. 547–552.

[19] R. Y. Lau, S. Y. Liao, R. C.-W. Kwok, K. Xu, Y. Xia, and Y. Li, "Text mining and probabilistic language modeling for online review spam detection," *ACM Trans. Manage. Inf. Syst.*, vol. 2, no. 4, pp. 1–30, 2011.

[20] F. H. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2011, vol. 22, no. 3, p. 2488.

[21] D. H. Fusilier, M. Montes-y-Gómez, P. Rosso, and R. G. Cabrera, "Detection of opinion spam with character n-grams," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Cham, Switzerland: Springer, Apr. 2015, pp. 285–294.

[22] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 309–319.

[23] M. Hazim, N. B. Anuar, M. F. A. Razak, and N. A. Abdullah, "Detecting opinion spams through supervised boosting approach," *PLoS ONE*, vol. 13, no. 6, 2018, Art. no. e0198884.

[24] N. Kumar, D. Venugopal, L. Qiu, and S. Kumar, "Detecting review manipulation on online platforms with hierarchical supervised learning," *J. Manage. Inf. Syst.*, vol. 35, no. 1, pp. 350–380, Jan. 2018.

[25] D. Zhang, L. Zhou, J. L. Kehoe, and I. Y. Kilic, "What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews," *J. Manage. Inf. Syst.*, vol. 33, no. 2, pp. 456–481, Apr. 2016.

[26] S. Ahmed and A. Danti, "Effective sentimental analysis and opinion mining of Web reviews using rule-based classifiers," in *Computational Intelligence in Data Mining*, vol. 1. New Delhi, India: Springer, 2016, pp. 171–179.

[27] Y. Lin, T. Zhu, H. Wu, J. Zhang, X. Wang, and A. Zhou, "Towards online anti-opinion spam: Spotting fake reviews from the review sequence," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2014, pp. 261–264.

[28] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a general rule for identifying deceptive opinion spam," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2014, pp. 1566–1576.

[29] H. Ge, J. Caverlee, and K. Lee, "Crowds, gigs, and super sellers: A measurement study of a supply-driven crowdsourcing marketplace," in *Proc. ICWSM*, 2015, pp. 120–129.

[30] S. Shojaee, A. Azman, M. Murad, N. Sharef, and N. Sulaiman, "A framework for fake review annotation," in *Proc. 17th UKSIM-AMSS Int. Conf. Modelling Simulation*, Mar. 2015, pp. 153–158.

[31] N. Hussain, H. Turab Mirza, G. Rasool, I. Hussain, and M. Kaleem, "Spam review detection techniques: A systematic literature review," *Appl. Sci.*, vol. 9, no. 5, p. 987, 2019.

[32] A. C. Pandey and D. S. Rajpoot, "Spam review detection using spiral cuckoo search clustering method," *Evol. Intell.*, vol. 12, no. 2, pp. 147–164, Jun. 2019.

[33] A. U. Akram, H. U. Khan, S. Iqbal, T. Iqbal, E. U. Munir, and M. Shafi, "Finding rotten eggs: A review spam detection model using diverse feature sets," *KSII Trans. Internet Inf. Syst.*, vol. 12, no. 10, pp. 5120–5142, 2018.

[34] R. Narayan, J. K. Rout, and S. K. Jena, "Review spam detection using opinion mining," in *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*. Singapore: Springer, 2018, pp. 273–279.

[35] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *Inf. Process. Manage.*, vol. 56, no. 4, pp. 1234–1244, Jul. 2019.

[36] R. Ghai, S. Kumar, and A. C. Pandey, "Spam detection using rating and review processing method," in *Smart Innovations in Communication and Computational Sciences*. Singapore: Springer, 2019, pp. 189–198.

[37] X. Wang, K. Liu, and J. Zhao, "Handling cold-start problem in review spam detection by jointly embedding texts and behaviors," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 366–376.

[38] J. K. Rout, A. Dalmia, K.-K.-R. Choo, S. Bakshi, and S. K. Jena, "Revisiting semi-supervised learning for online deceptive review detection," *IEEE Access*, vol. 5, pp. 1319–1327, 2017.

[39] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, vol. 57, pp. 117–126, Sep. 2016.

[40] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Inf. Sci.*, vol. 181, no. 6, pp. 1138–1152, Mar. 2011.

[41] S. Krishnamoorthy, "Linguistic features for review helpfulness prediction," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3751–3759, May 2015.

[42] Y. Dang, Y. Zhang, and H. Chen, "A lexicon-enhanced method for sentiment classification: An experiment on online product reviews," *IEEE Intell. Syst.*, vol. 25, no. 4, pp. 46–53, Jul. 2010.

[43] R. Moraes, J. F. Valiati, and W. P. Gavião Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Syst. Appl.*, vol. 40, no. 2, pp. 621–633, Feb. 2013.

[44] G. Vinodhini and R. M. Chandrasekaran, "Opinion mining using principal component analysis based ensemble model for e-commerce application," *CSI Trans. ICT*, vol. 2, no. 3, pp. 169–179, Nov. 2014.

NAVEED HUSSAIN received the M.S. degree in computer science from the University of Central Punjab, Lahore, Pakistan. He is currently pursuing the Ph.D. degree in computer science with COMSATS University Islamabad, Lahore Campus, Pakistan. He is also an Assistant Professor with the Department of Software Engineering, The University of Lahore, Lahore. He has published several articles in reputed journals. His research interests include data mining, sentimental analysis, machine learning, and opinion mining.

HAMID TURAB MIRZA received the Ph.D. degree in computer science from Zhejiang University China, in 2012, and the M.Sc. degree (Hons.) in information systems from The University of Sheffield, England, in 2005. He has more than 15 years of research, teaching, and information systems development experience in aviation, telecommunication, and academic sectors of both Pakistan and U.K. His research interests include within the areas of data mining, machine learning, and human–computer interaction. He has published several articles at reputed international conferences and journals in these areas.

IBRAR HUSSAIN received the M.S. degree in information management (information retrieval) from the Queen Mary University of London, U.K., in 2006, and the Ph.D. degree in computer science from Zhejiang University, China, in 2014. He is currently an Associate Professor and the Head of the Department of Software Engineering, The University of Lahore, where he has been, since 2015. His research interests include within the areas of human–computer interaction, artificial intelligence, machine learning, ubiquitous computing, accessibility, and location-based services.

FAIZA IQBAL received the M.S. and Ph.D. degrees from the National University of Sciences and Technology (NUST), Islamabad. She is currently working as an Assistant Professor with The University of Lahore, Lahore. She has also been with the Department of Computer Science, Quaid-i-Azam University, Islamabad. She has published several articles in reputed journals and conferences. Her areas of research include machine learning, data analysis of the IoT Systems and high-performance network optimization.

IMRAN MEMON received the B.S. degree in electronics from the IICT University of Sindh, Jamshoro, Pakistan, in 2008, and the M.E. degree in computer engineering from the University of Electronic Science and Technology, Chengdu, China. He is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Zhejiang University. He is also an Assistant Professor with the Department of Computer Science, The Bahira University, Karachi Campus. He has published over 30 research articles in recent years. He serves as an organizing committee Chair and a TPC member more than 200 international conferences, as well as a reviewer for over 50 international research journals. He also serves as an Editor-in-Chief of *Journal of Network Computing and Applications*.