Finding Rotten Eggs: A Review Spam Detection Model using Diverse Feature Sets

[1]Abubakker Usman Akram, [1]Hikmat Ullah Khan, [2]Saqib Iqbal, [1,*] Tassawar Iqbal, [1]Ehsan Ullah Munir, [3]Dr. Muhammad Shafi

[1]Department of Computer Science, COMSATS Institute of Information Technology, Wah Cantt, Pakistan
[2]Department of Software Engineering, Al-Ain University of Science and Technology, Al-Ain, UAE
[3]Department of Computer Science, Air University, Islamabad, Pakistan
[Email: abubkr.akram@gmail.com, hikmat.ullah@ciitwah.edu.pk, saqib.iqbal@aau.ac.ae, tassawar@ciitwah.edu.pk, ehsan@ciitwah.edu.pk, mshafi@mail.au.edu.pk]
*Corresponding Author: Tassawar Iqbal

Received February 2, 2018; revised April 12, 2018; accepted April 23, 2018; published October 31, 2018

Abstract

Social media enables customers to share their views, opinions and experiences as product reviews. These product reviews facilitate customers in buying quality products. Due to the significance of online reviews, fake reviews, commonly known as spam reviews are generated to mislead the potential customers in decision-making. To cater this issue, review spam detection has become an active research area. Existing studies carried out for review spam detection have exploited feature engineering approach; however limited number of features are considered. This paper proposes a Feature-Centric Model for Review Spam Detection (FMRSD) to detect spam reviews. The proposed model examines a wide range of feature sets including ratings, sentiments, content, and users. The experimentation reveals that the proposed technique outperforms the baseline and provides better results.

Keywords: Review spam detection, feature engineering, sentiment analysis, rating diversity, social networks

http://doi.org/10.3837/tiis.2018.10.026 **ISSN : 1976-7277**

1. Introduction

Social media enables people to generate and share content. This content generation facility provides a huge information repository, but raises many problems such as munging, advance-fee scam, identity-theft, phishing attacks [1], spam [2], email fraud, and mobile malwares. Currently, spam is one of the main concerns which the people are facing. Spam is an unwanted electronic information spread by the spammers with an aim to cause monetary and psychological damage to the victims [3].

Spam can be categorized into numerous types [4, 5], but the most common are email spam, SMS spam, web spam, and review spam. Mail spam is related to unwanted electronic messages [6, 7]. According to a recent study[1], nearly 56.87% emails are classified as spam. SMS spam is unwanted messages that are delivered to customers, and are not only annoying but may cause financial loss to the service providers [6]. Web spam is usually used to deceive search engines to make wrong decisions in ranking of web pages [8].

Review spam is a growing problem in which the spammers often exploit reviews by giving wrong or false positive reviews. The review spam usually targets both customer and companies. In the former case, it may misguide customers to make wrong decisions about a product, whereas in the later case, the review spam may result in huge loss for companies [9]. The review spam can be further classified into three categories, namely untruthful reviews, reviews on brands, and non-reviews [10]. The untruthful reviews are false positive or false negative in nature. The aim of such reviews is to falsely promote a certain product or to damage reputation of the competitors.

A review on brands entails the brand-based personal experience or knowledge instead of a specific product. Although this category of review spam is helpful for judging a brand, doing so with the perspective of a single product is not realistic. A non-review is a spam review, which does not contain any positive information at all and promotes advertisements or random content. Considering the severity of the problem, there is a need for an effective spam detection technique that is applicable to a wide range of spam categories.

In this paper, we propose a review spam detection model named Feature-centric Model for Review Spam Detection (FMRSD). Our main research contributions are as under:

- An effective model is proposed to detect review spam based on links, users, rating, sentiments and content based features.
- A set of novel features is proposed and applied on Amazon Dataset[2].
- Sentiment analysis is performed considering the prestige of a user.

[1] https://www.statista.com/statistics/420391/spam-email-traffic-share/ Accessed on: 12/July/2017
[2] https://snap.stanford.edu/data/web-Amazon.html/ Accessed on: 22/October/2016

- Diverse feature sets are examined to avoid false spam detection.

The rest of the paper is organized as follows. Section 2 reviews the related work, Section 3 presents the proposed model, Section 4 defines the experimental setup, Section 5 discusses results, and Section 6 concludes the paper.

2. Related Work

The exponential growth of the social web in which users generate their own content has resulted into diverse research problems, such as spam-detection, phishing and malware detection, online fraud detection, and reputation and trust management [2, 11, 12]. Due to various types of spams, such as web spam, review spam, email spam etc., the spam-detection process is a challenging task [13, 14]. Web spam is a common type of spam which is used to deceive search engines to falsely rank web pages and this type of spam is usually detected using link-based approaches [15] and graph based PageRank algorithms [16], [17].

In PageRank algorithms, it is assumed that the pages having a varied distribution of the linked pages are considered as suspicious page. Thus such suspicious pages having a false high PageRank are classified as spam. In general, this assumption may not be true as a spammer may have some genuine links and will have lower chances to be detected. Similarly, graph based approaches are also used for review spam detection [11], however, they ignore the review content. To overcome this shortcoming, Li *et al.*, [18] used graph-based approach that considers content. This approach takes a sentence as a node, assigns a weight to each node according to its significance based on probability of being spam, and mark the nodes below the set threshold as spam. Although this technique is better compared to the prior one, it does not consider the prestige of users and a new user with little domain knowledge may be wrongly detected as a spammer [19].

Supervised learning algorithms are also used for review spam detection. Jindal and Lie [20] used Naïve Bayes classifier to detect spam and found that classification algorithms are not good for spam detection. Moreover, they proposed an extension by introducing new features to detect spam, and found improved results [21]. Lim *et al.*, [22] proposed a method of supervised learning that considers users' review ratings. The proposed method showed promising results and was helpful in identifying spammers and detecting review spam. In addition to finding individual spammer, Mukherjee *et al.*, [23] proposed a method to detect group spammers by analyzing group behaviour.

In this regard, many researchers have also used Natural Language Processing (NLP) technique which exploits content based features to detect the unusual patterns in the content. In [24], the authors proposed a content-based approach using lexical and sentiment features. According to the authors' assumption, if a review contains too positive or too negative sentiments, then its most probably spam and must be flagged as spam. However, the assumption is not a good measure as a customer may be completely satisfied with a product and may have written a valid review. Consequently, the false positive rate of this technique will be higher. To avoid such limitations of the content based techniques, Yuming *et al.*, [25] proposed to consider users' review behavior. To detect the review spam, link-based and network-based techniques are also used. In link-based spam, one of the techniques used by the spammers is cloaking which feeds spam content to the search engines. In [26], the authors proposed a link-based technique to detect link-based cloak spam. The authors used a modified version of PageRank algorithm to detect the users' behavior and subsequently detect the spam web pages. Link-based techniques are more useful to detect web spams as search engines use the inlinks and outlinks to index pages. To check the authenticity of inlinks and outlinks, Junting *et al.*, [27] proposed a network-based approach which uses 2-hop sub-graphs to top ranking products in online reviews. The method is feasible for review spam detection of old products, but the false positive rate may increase for new products as the methodology fails to consider the content features.

Another approach considers temporal features which use time to detect the review spam. In [28], the authors used temporal approach to detect spams and is based on the burst patterns and frequency of reviews. Chen and Chen [29] used similar burst patterns approach of reviews along with a temporal feature to detect the behaviour of the users. A burst pattern may not be a good indicator to judge the spam, as a user may be writing reviews for many products. In [30], the authors used temporal based rating-consistency to detect spam and marked abnormally frequent reviewers as spammers. In [31], the authors criticized the temporal approach and considered it unuseful as spammer may post fewer reviews. Consequently, these reviews may remain undetected even if the content of the reviews is spam. In [32], the authors proposed a temporal and spatial feature based technique to overcome the limitations of the temporal factors and found improved results.

The link or content-based approaches alone are not considered sufficient as spammers usually make their reviews look normal so they can deceive spam detection methods easily. In this regard, hybrid approaches are used to detech spam which consider both link and content of a review. In [33], the authors used a hybrid approach to detect spam using both link-based and content-based techniques. They used a graph topology to find neighbours and later applied the content-based technique of majority voting mechanism to detect spam. Another hybrid approach based on content and network features was proposed by Rayana and Akoglu [32]. As the approach does not consider behavioural features, thus rarely effective to detect spams when the user is not habitual. In [34], the authors proposed another hybrid approach to detect review spams that considers behaviour and user features. However, the approach is not feasible to detect content-based spams as behaviour and user features are insufficient to detect it [10]. In [35], the authors proposed a trust index technique to detect review spams. The technique uses an iterative algorithm along with content features, but lacks user features. In [36], the authors used used both behavioual and linguistic features and found that behavioural features outperforms the linguistic features.

It is evident from the literature review that single feature sets are not effective against review spam detection [33], and hybrid approaches perform better. Consequently, there is a need for a new method that can consider diverse feature sets and detect review spam efficiently.

3. Review Spam Detection Model

In order to detect review spam, we propose a Model named Feature-centeric Model for Review Spam Detection (FMRSD). The proposed framework consists of four phases as shown in **Fig. 1**. In the first phase of preprocessing, data cleansing is performed to remove the noisy data, and cleaned data is stored in the database for further processing. In the second phase, six diverse feature sets are computed using the techniques explained in the Section 3.1. In the third phase of data transformation, all the features are normalized in a range of 0-1 using min-max normalization technique. Finally, normalized features are fed to the proposed algorithm to detect whether the reviews are spam or not. During the final phase, Kendall's correlation, Spearman's correlation and Osim are used to evaluate the results. All the six divers features sets, exploited in the porposed framework, are discussed in detail in the next sub-section. **Table 1** represents the list of symbols used in the paper.

Table 1. List of Symbols used in the paper

U	Set of Users.
P	Set of Products
R	Set of Reviews
F	Feature
FS	Feature Set
N	Number
T	Time
S	Score
W	Word in a review
N_{UR}	Number of Reviews by a User
N_{UAD}	Number of Active Days of a user
S_{RH}	Helpfulness Score of a Review
N_{PV}	Number of Votes given to a Product
N_{PR}	Number of Reviews given to a Product
N_{RW}	Number of Words in a Review
$\overline{N_{PW}}$	Mean Number of Words per review of a Product

S_{RS}	Combined Sentiment Score of a Review
S_{WS}	Combined Sentiment Score of a Word
S_{WPS}	Positive Sentiment Score of a Word
S_{WNS}	Negative Sentiment Score of a Word
S_{WNeuS}	Neutral Sentiments Score of a Word
$\overline{S_{PS}}$	Mean Sentiment Score of a Product
S_{UR}	Rating Score given by a User
$\overline{S_{PR}}$	Mean Rating Score of a Product
T_{PL}	Launch Time of a Product
T_{PR}	Review Time of a Product
N_{RFP}	Number of First Person Pronouns in a Review
N_{RSP}	Number of Second Person Pronouns in a Review
N_{RTP}	Number of Third Person Pronouns in a Review

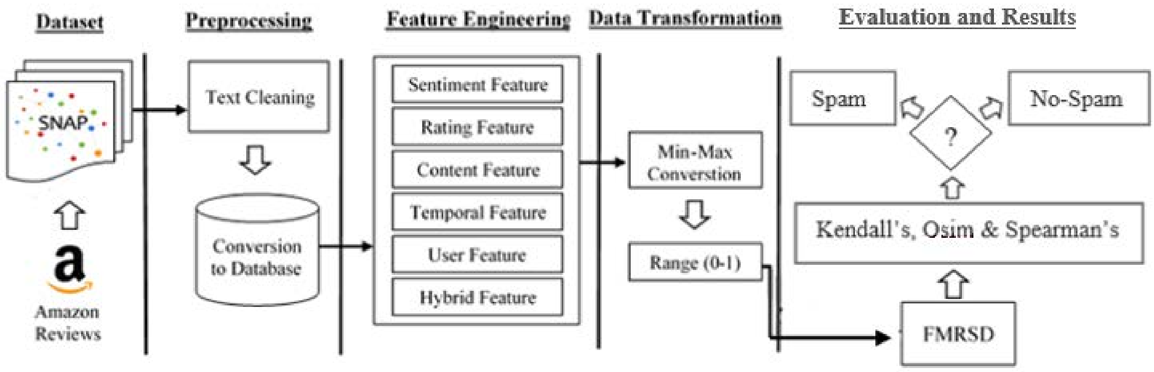

Fig. 1. The Proposed Spam Detection Framework

3.1 Features Engineering

The feature sets are classified into six categories, namely Rating-based, User-Based, Temporal-based, Sentiment-based, Content-based, and Hybrid. Moreover, we have merged rating and sentiment features to compute hybrid feature set because hybrid approaches show promising results.

In Rating-based feature, review ratings are used by users to rank a product, however, spammers often rate a product either too good or too bad which tends to deviate from the mean value. If the difference from mean rating is greater than 2/3, then the review is considered to be a spam. To do so, firstly the product mean rating score $\overline{S_{PR}}$ is calculated using the equation (1):

$$\overline{S_{PR}} = \frac{\sum_{\mathrm{i}=1}^{N_{PR}} S_{UR}}{N_{PR}} \tag{1}$$

Secondly, equation (2) is used to find deviation of the rating (F_{RD}) for ith review having rating X.

$$F_{RD_{\mathbf{i}}} = |\mathrm{X_i} - \overline{S_{PR}}| \tag{2}$$

The rating-based technique is insufficient to detect a review spam because if the rating falls within the mean threshold, then the review will not be flagged as spam. Hence, the reviewing frequency(F_{URF}) is an important characteristic that confirms the legitimance of a user [37]. Moreover, the process of reviewing takes much time, whereas spammers review many products in a short span of time. Thus, to compute reviewing frequency (F_{URF}), equation (3) is used:

$$F_{URF} = \frac{\sum_{\mathrm{i}=1}^{N_{UR}} \mathrm{R_i}}{(\mathrm{N_{UAD}}/7)} \tag{3}$$

Where, number of reviews are calculated by adding all reviews (N_{UR}) by a user in all active days. To compute the number of reviews of a reviewer per week on average, the number of reviews of a reviewer are divided by ratio of number of active days of a user ($\mathrm{N_{UAD}}$) to number of days in a week. If $F_{URF} \geq 10$, then a reviewer is considered as a spammer.

The frequency of reviews is only applicable, if the user is a regular spammer. However, if a reviewer reviews a product soon after it is launched, then review will more likely be a spam. Therefore, we compute the total time (F_{TR}) between the product listing time and the review time using equation (4). As the frequency of review rating is a a boolean feature, so if a product is reviewed within three days, then we classify it as a spam.

$$F_{TR} = T_{PL} - T_{PR} \tag{4}$$

$$If\,(F_{TR} \leq 3)\; F_{TR} = \mathbf{1}$$
$$else\; F_{TR} = \mathbf{0}$$

A review may be positive or negative. A neutral review is a sign that the review does not explain anything good or bad about the product. We have used SentiWordNet [38] for opinion mining which is widely used lexical resource for opinion analysis [39]. The sentiment score computed by the lexicon is from -1 to +1 with 0 in middle to represent neutral review. The value less than 0 represents negativity in the review, so lower the value the higher will be the negative sentiments in the review. Similarly, the value greater than 0 represents positive sentiments and higher value represent higher positivity in a review. If the review is neutral i.e., having 0 output value, then it is more likely to be a spam containing useless information. We first compute review sentiment score by adding the sentiment score of each word in a review, represented by (S_{WNeuS}). Then to get the normalized value of the review neutrality (F_{RN}), we devide S_{WNeuS} by number of words. The neutrality of the review is calculated by equation (5).

$$F_{RN} = \frac{\sum_{i=1}^{N_{RW}} S_{WNeuS_i}}{N_{RW}} \tag{5}$$

A product may consist of some likeable and unlikable features based on personal preferences. If a review is highly positive or too negative, it might be a spam [40]. Highly positive review means the reviewer is exaggerating product features and too negative review shows biaseness against a product. The positive sentiment ranges from 0 to +1 and negative seniment ranges from 0 to -1. If the difference of positive and negative seniments in a review exceeds 2/3 of the accumulative sentiment score of the total words present in a review content, then it means the review is eighter too positive or too negative. A review in both cases can be a spam. The review sentiment difference F_{RSD} is calculated using equation (6).

$$F_{RSD} = \left|\sum_{i=1}^{N_{RW}} (S_{WPS_i} - S_{WNS_i})\right| \tag{6}$$

Moreover, as majority of reviewers have similar views about a certain product, if there is a high sentiment deviation (F_{SD}) between the mean sentiment score ($\overline{S_{PS}}$) and the sentiment score (S_{WS}) for a specific review, it is likely to be a spam [21]. $\overline{S_{PS}}$ is calculated using equation (7) and F_{SD} is calculated using equation (8).

$$\overline{S_{PS}} = \frac{\sum_{i=1}^{N_{RW}} S_{WS_i}}{N_{RW}} \tag{7}$$

$$F_{SD} = \left||S_{RS}| - \overline{S_{PS}}\right| \tag{8}$$

A review should also define self-experience of a user about a product. If there are too many 2nd or 3rd person pronouns in a review, then it might be a spam; as it is sharing or commenting others views. We have used the ratio of 2nd and 3rd person pronouns with the 1st person (F_{CSE}) using equation (9).

$$F_{CSE} = \frac{N_{RSP} + N_{RTP}}{N_{RFP}} \tag{9}$$

Some products are small and less complex, such as wrist bands, and require a small review, whereas some are complex, such as laptops, and they need a detailed review. If the review is too small or too long F_{CRL} as compared to mean length of reviews $\overline{N_{PW}}$ for a certain product, then it is considered as a spam [41]. $\overline{N_{PW}}$ is calculated using equation (10).

$$\overline{N_{PW}} = \frac{\sum_{i=1}^{N_{PR}} N_{RW_i}}{N_{PR}} \qquad (10)$$

Moreover, F_{CRL_i} is presented using eq. (11) for the ith review having N_{RW} words.

$$F_{CRL_i} = \left| N_{RW_i} - \overline{N_{PW}} \right| \qquad (11)$$

Spammers often rate products differently as compared to their comments in the review. For instance, if a reviewer rates a product good and provides negative comment in the review of a product, then it is likely to be a spam. To calculate the ratio of review rating and sentiment score of a product, F_{HRS} for ith review is computed using equation (12).

$$F_{HRS_i} = \left| {}^{X_i}/_{5} - |S_{RS}| \right| \qquad (12)$$

Where, the normalized rating score in the range 0-1 is obtained by dividing rating by 5 as rating is in the range of 1-5. The higher the F_{HRS_i} score, the higher will be the difference so the higher chances of a spam review.

3.2 The Proposed Algorithm

The proposed model FMRSD, is described through the following algorithm. The algorithm takes the reviews data as input and computes the feature sets. The time and space complexity for the computation of all the feature sets is linear, i.e., O(n). If the resultant rank is less than the threshold rank of 40, the review is considered to be normal else it will be considered as spam [42].

FMRSD Algorithm:

Input: Data of Amazon reviews
Output: Review detection as spam or no spam.

```
1.  For each review r ∈ R, in a product p ∈ P, by a user u ∈ U.
    2.  Initialize $N_{PV}$, $N_{PR}$ , $N_{RW}$
    3.  $N_{PV}$ = countProductVotes (p)
    4.  $N_{PR}$ = countProductReviews (p)
    5.  $N_{RW}$ = countReviewWords (r)
        ▷ Computation of Rating Feature Set ($FS_{RB}$)
    $S_{PR} =$ computeSumofRatings (p)
    $X_i$ = retriveRatingofUser (r)
    $\overline{S_{PR}} = S_{PR} / \overline{N_{PR}}$
    $F_{RD} = |X_i - \overline{S_{PR}}|$
    $FS_{RB} = [FS_{RB}; F_{RD}]$
    ⊳ **Computation of User Feature Set ($FS_{UB}$)**
    $N_{UR}$ = countNumberofRatings (u)
    $F_{URF} = N_{UR}/(N_{UAD}/7)$
    $FS_{UB} = [FS_{UB}; F_{URF}]$
    ⊳ **Computation of Temporal Feature Set ($FS_{TB}$)**
    $T_{PL}$ = fetchProductLaunchTime (p)
    $T_{PR}$ = fetchProductReviewTime (r)
    $F_{TR} = T_{PL} - T_{PR}$
    $FS_{TB} = [FS_{TB}; F_{TR}]$
    ⊳ **Computation of Sentiment Feature Set ($FS_{SB}$)**
    $\overline{S_{PS}}$ = computeMeanSentimentScore (p)
    $S_{RS} =$ ComputeSentimentScore(r)
    $Sum_{RS}$ = SumCombinedSentimentScore (r)
    $Sum_{RPS}$= SumPostiveSentimentScore (r)
    $Sum_{RNS}$= SumNegativeSentimentScore (r)
    $Sum_{RNeuS} =$ SumNeutralSentiments(r)
    $F_{RN} = \frac{Sum_{RNeuS}}{N_{RW}}$
    $F_{RSD} = |Sum_{RPS} - Sum_{RNS}|$
    $F_{SD} = \left| |S_{RS}| - \overline{S_{PS}} \right|$
    $FS_{SB} = [FS_{SB}; F_{RN}, F_{RSD}, F_{SD}]$
    ⊳ **Computation of Content Feature Set ($FS_{CB}$)**
    $N_{RFP}$ = CountFirstPersonPronouns (r)
    $N_{RSP}$ = CountSecondPersonPronouns (r)
    $\underline{N_{RTP}}$ = CountThirdPersonPronouns (r)
    $\overline{N_{PW}}$= ComputeProductMeanLength (p)
    $F_{CSE} = (N_{RSP} + \underline{N_{RTP}})/ N_{RFP}$
    $F_{CRL} = |N_{RW} - \overline{N_{PW}}|$
    $FS_{CB} = [FS_{CB}; F_{CSE}, F_{CRL}]$
    ⊳ **Computation of Hybrid Feature Set ($FS_{HB}$)**
    x = RetriveRating (r)
    $S_{RS}$ = computeSumofSentimentScore (r)
    $F_{HRS_i} = \left| {}^{x_i}/_{5} - |S_{RS}| \right|$
    $FS_{HB} = [FS_{HB}; F_{HRS}]$
End For
Rank = FMRSD ($FS_{RB}, FS_{UB}, FS_{TB}, FS_{SB}, FS_{CB}, FS_{HB}$ )
IF Rank >= 40 THEN Spam
```

```
42.  Else NoSpam
43.  STOP (END of Algorithm)
```

4. Experimental Setup

The following sub-sections present the experimental setup for both baseline and the selected dataset. Afterwards, the performance evaluation measures are discussed, which show the state of the art methods for examining the authenticity of the proposed technique.

4.1 Baseline

The adapted version of PageRank with content weights, which is used to identify spam in web pages, is taken as a baseline [43]. The PageRank score r(p) of a page p is defined as :

$$r(p) = \alpha \sum q:(p,q) \in \varepsilon \frac{r(q)}{\omega(q)} + (1-\alpha)\frac{1}{N} \tag{14}$$

Where, α is a decay factor, $r(q)$ is PageRank of a page q that is out linked by p and $\omega(q)$ is the out degree of q.

4.2 Dataset

For the experiments, a dataset of the SNAP Stanford repository is used [39, 40]. Using custom built software, the data was cleaned and stored in a MySQL database for further processing. The statistics of the dataset are shown in **Table 2**. Ratings given for the products are in the range from 1 to 5, where 1 represents lower rating and 5 represents higher rating. The dataset contains both spam and not-spam reviews.

Table 2. The Dataset Statistics

Reviews	34,686,770
Users	6,643,669
Products	2,441,053
Avg. Number of Reviews per Product	~ 14
Users with > 50 reviews	56,772
Median no. of words per review	82
Timespan	Jun 1995 - Mar 2013

4.3 Performance Evaluation Measures (PEM)

This section describes performance evaluation measures of the results. These measures include, two types of correlations and Osim. Details are provided in the following sections.

4.3.1 Kendal's Rank Correlation

The first measure, Kendal's Rank Correlation [44] is used to calculate the rank correlation between features calculated in this study and values of baseline. It also compares the variation between features calculated and baseline values. A Pair of the result is said to be concordant, if the pair of result increases along other pair in the data. Otherwise, it is considered discordant. Kendal's Rank Correlation is represented by τ and calculated using equation (15).

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}k(k-1)} \tag{15}$$

4.3.2 Spearman's Rank-Order Correlation

The second measure i.e., Spearman's Rank-Order correlation is used to compute the correlation between two rank orders [45]. It is represented by ρ and calculated using equation (16).

$$\rho = \frac{n(\sum R_1 R_2) - (\sum R_1)(\sum R_2)}{\sqrt{[k\sum {R_1}^2 - (\sum R_1)^2][k\sum {R_2}^2 - (\sum R_2)^2]}} \tag{16}$$

Where, $\mathbf{R_1}$ and $\mathbf{R_2}$ represent the results of the proposed method and baseline respectively

4.3.3 Osim

The third measure i.e., Osim [46] is used to calculate the intersection between the pair of values from $\mathbf{R_1}$ and $\mathbf{R_2}$ and is represented by equation (17).

$$\text{OSim} = \frac{R_1 \cap R_2}{k} \tag{17}$$

Here, k is the number of records in the list on which correlation is calculated.

5. Results and Discussions

The proposed and the baseline techniques are correlated using Kendal's correlation, and results are shown in **Table 3**. The results show that the proposed FMRSD has outperformed the baseline technique. The Spearman's correlation of baseline technique with the helpfulness is 0.472. In addition, the feature sets of the proposed technique also showed promising results compared to the existing approach. For instance, the sentiment feature set that examines the diversity in sentiments along with the neutrality in the review has a score of 0.568. Similarly, the user feature set that tests the behaviour of the user has a score of 0.474. Moreover, the hybrid feature set consisting of both sentiment and rating diversity has a score of 0.563. Importantly, FMRSD correlation score 0.594 is superior to the scores of all feature sets.

Table 3. Kendall's Correlation

Method	**Helpfulness**
Adopted PageRank	0.472
Rating feature set	0.481
User feature set	0.474
Temporal feature set	0.478
Sentiment feature set	0.568
Content feature set	0.547
Hybrid feature set	0.563
FMRSD	**0.594**

The sentiment feature set is comprised of three sub features, i.e. review neutrality, high valence score, and sentiment deviation from mean sentiment score of a product. A comparison of the baseline with the sentiment feature set and its sub features is shown in **Fig. 2**. The review neutrality has a high score of 0.437 which confirms the presence of spam reviw.Moreover, the high valence has scored 0.334 which is less than the other sub-feature set because many customers buy products just to satisfy their specific needs [10]. Usually, normal customers are not critics in nature and define the general working of the product. Examining only the positive or negative sentiments is important for the presence of spam, but only when they have a high false positive rate. The results show that the sentiment deviation has a score of 0.397 and performed relatively much better than the high valence feature. The sentiment deviation is based on the difference between the sentiment score of a review and mean sentiment score for a product. If a reviewer has shown experience with the product in contrast with all other reviews, its more likely to be a spam. The score is relatively low as compared to review neutrality because of false positive rate. A customer may face a problem with a product unlike other reviewers, for instance, a bought computer may have a faulty hard disk and it may fail immediately. Likewise, it is also observed that a product may satisfy few users. These rare likes/defects in a product may lead to a different review from a normal one.

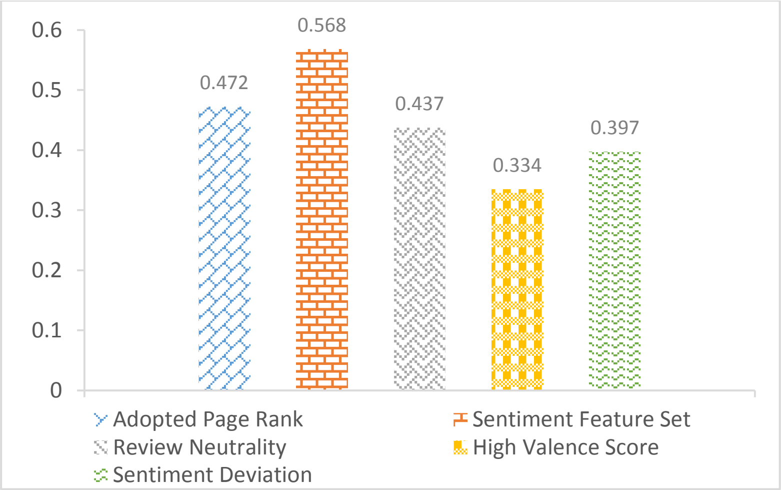

Fig. 2. Kendall's Correlation with Sentiment based sub-features

The content of a review is highly important in review spam evaluation [47]. The content feature set comprises of two sub-features, namely self-experience and review-length ratio. The results of the content sub-features are shown in **Fig. 3**. In self-experience, the presence of singular pronoun like "I" and "me" is checked against third person plural pronouns like "they" and "their". The presence of more third person plural pronouns show that the reviewers are not sharing their self-experience, but just mentioning others experiences in the review. The presence of first singular pronouns in a review indicates the evidence of self-experience, mentioned in a review. This may have a false positive rate as a person may be defining the experience of others with the product, for instance, a husband may have bought some product for his wife and just defining her experience with the product. As this scenario of explaining others experience is rare so the score is 0.42, which is a valuable contribution to the total score of the content feature set. The results show that the review length ratio has scored significantly less than the self-experience because of high false positive rate. Although, the short review is a good measure to check spam, many genuine reviews may be short. A short review may be a result of the busy daily life of customer. A customer may have liked anything or disliked a product because of a particular reason, for instance, a customer may have bought a new laptop for prime battery time and may have reviewed just based on the observed battery time for the laptop.

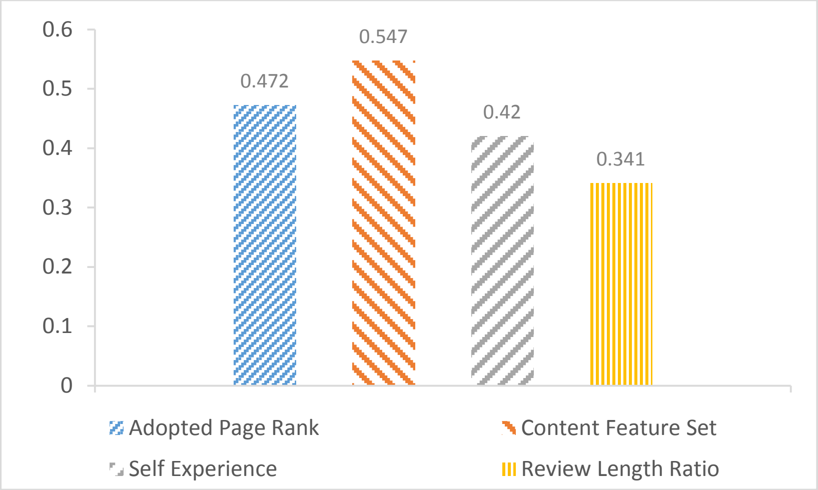

Fig. 3. Kendall’s Correlation with Content based sub-features

Table 4 presents Spearman correlations of features along helpfulness. It shows that the proposed FMRSD with the top score of 0.62 has outperformed the baseline technique with a score of 0.521.

Table 4. Spearman's Rank Order Correlation

Method	**Helpfulness**
Adopted PageRank	0.521
Rating feature set	0.527
User feature set	0.526
Temporal feature set	0.533
Sentiment feature set	0.611
Content feature set	0.562
Hybrid feature set	0.559
FMRSD	**0.621**

FMRSD is capable of generating good results because of the quality of the features. Moreover, the rating feature set which isbased on rating deviation is found effective in detecting spams because of the nature of the features. Similarly, temporal and user feature sets have performed slightly better than the baseline technique. As the baseline is link-based only, it is less effective for web spam detection.

The sentiment features set has provided a much better score because the feature set not only checks the intent of a reviewer, but also checks the presence of information in it. For instance, the sentiment feature set has a score of 0.611 and the content feature set has a score of 0.562.

The sentiment feature set presents promising results because of powerful sub-features, such as review neutrality, sentiment deviation, sentiment featue set, and high valence score, shown in **Fig. 4**. The results show that the review neutrality has a score of 0.447 and is the highest among all other sub features. The reason for a high score is less false-positive and large true-positive cases. The presence of no sentiments in a review leads to the neutrality of the review and thus is a proof of spam. The high valence score is based on too positive or too negative review and has a comparatively high false-positive rate. According to the results, the sentiment deviation has a score of 0.402 and performed much better than the high valence score because of the less false-positive rate. Moreover, the sentiment deviation is a good measure to detect a non-habitual spammer. The habitual spammers are normally smart enough to blend the review with some sentiments to avoid from being detected as spam.

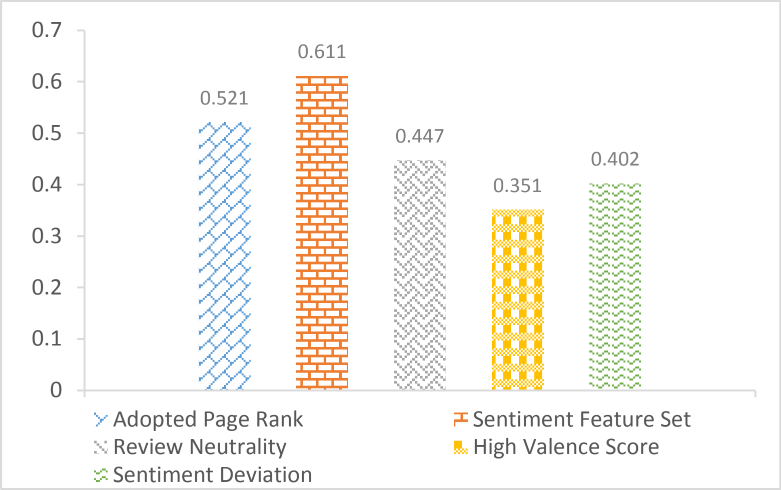

Fig. 4. Spearman's Correlation with Sentiment based sub-features

As discussed earlier, the content-features are based on self-experience and review-length ratio. The results show that the self-experience has scored 0.437, which is a significant contribution to the total of the content-feature set as shown in **Fig. 5**. The self-experience has scored 0.437, which is much better than the review length ratio score of 0.35. The reason for a low score for the review-length ratio is the increase in the number of false-positive cases, as short reviews are not always spam. When these two features are merged, a significant increase is found in the score, i.e., 0.562. The reason behind the decreased false-positive and increased true-positive rate is that a review which is short but defines self-experience may have been missed from false positive results by combining the two features. Likewise, a review written on behalf of others may avoid the wrong-detection due to acceptable length of the review.

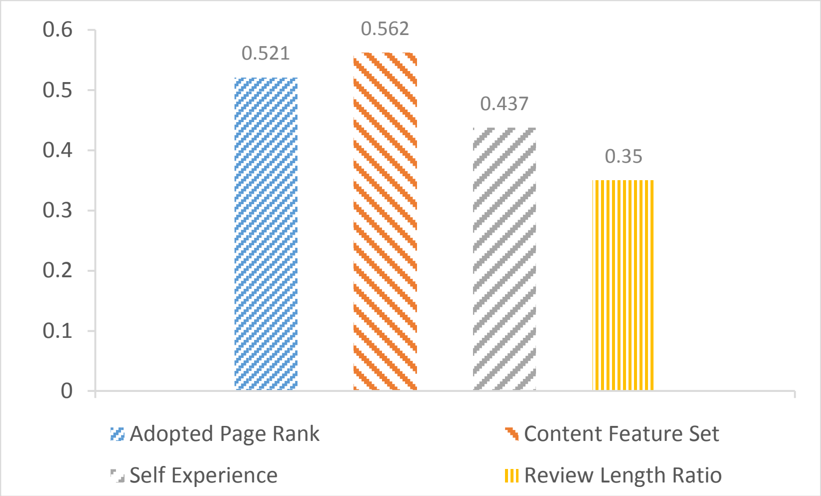

Fig. 5. Spearman's Correlation with Content based sub-features

The content-features based on self-experience and other feature sets have produced promising results as evident by Osim values shown in **Table 5**. According to the results, the baseline technique has scored 0.517.

Table 5. Osim values for diverse Feature sets

Method	**Helpfulness**
Adopted PageRank	0.517
Rating feature set	0.523
User feature set	0.522
Temporal feature set	0.534
Sentiment feature set	0.582
Content feature set	0.564
Hybrid feature set	0.536
FMRSD	**0.613**

The results also show that individual features have scored better. As the baseline technique totally relies on the number of links, it completely ignores the content and behaviour of the users. The proposed FMRSD not only considers the content and user behaviour, but also have other novel feature sets which help to detect correct review spam. As FMRSD is capable of avoiding false spam detection, the false positive rate is much lower than the baseline technique. The feature sets have also performed significantly better than the baseline approach. For instance, the sentiment feature set has scored 0.582, which is better than the baseline 0.517. Similarly, the results show that the score of content features is 0.564, which is also better than the baseline technique. It is evident that the individual feature sets have shown better results.

The sentiment-feature set is comprised of sub-features, namely review neutrality, sentiment deviation, sentiment feature set, and high valence score as shown in **Fig. 6**. The results show that the review-neutrality is an important sub-feature having score of 0.441, and it is because of lower false positive rate. Moreover, the high valence score of 0.323 and sentiment deviation of 0.387 performed less than the review neutrality sub-feature, but when combined, it performed significantly well due to the decreased false-positive rate. Thus, a reviewer that has a neutral review or fewer sentiments may be real, and may have not been detected as a spam due to high valence score.

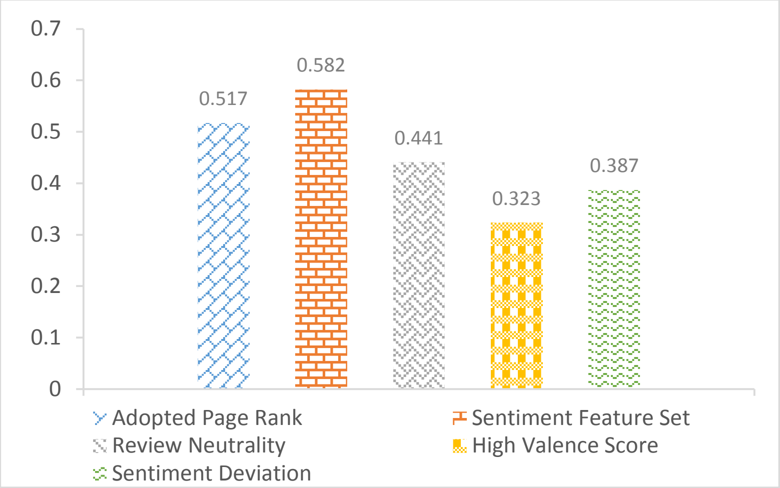

Fig. 6. Osim values with Sentiment based sub-features

The content-based feature results and sub-feature results are shown in **Fig. 7**.

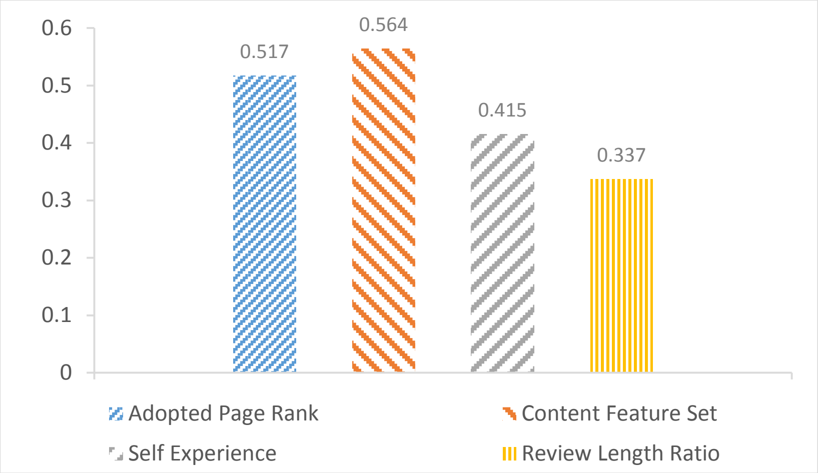

Fig. 7. Osim values with Content based sub-features

The results show that self-experience and review-length ratio does not perform as good as other sub-features because of the low true-positive rate. Moreover, when the sub-features are merged, they perform significantly better than the baseline technique. The false-positive rate is lowered by combining the sub-features that increases the overall score of Osim. Also, the self experience score of 0.415 is significant, but cannot represent the presence or absence of spam.

6. Conclusion

In this paper, a hybrid technique named FMRSD is proposed to detect spam reviews. The proposed technique is comprised of several feature sets. The results confirmed that the proposed technique outperforms the baseline method. More specifically, the sentiment-based features have the highest correlation with helpfulness as compared to the rest of the feature sets. Also, the hybrid of sentiment and rating features has a relatively better correlation. The content features are also proved significant as they consider the characteristics of the content along withopinion or sentiments expressed in the review. Moreover, the significance of individual feature in each feature set is also determined. For sub-category, sentiment-neutrality showed optimal results, whereas the sentiment-deviation from the mean value is more significant as compared with the sentiment-valence. Among the content-based features, the lexical features are more significant compared to the review length. In future, our aim is to use the proposed model for detection of social bots.

References

[1] Dadkhah, M., et al., "An overview of phishing attacks and their detection techniques," *International Journal of Internet Protocol Technology*, 9(4), p. 187-195, 2016. Article (CrossRef Link).

[2] Khan, H.U., et al., "Modelling to identify influential bloggers in the blogosphere: A survey," *Computers in Human Behavior*, 68, p. 64-82, 2017. Article (CrossRef Link).

[3] Shen, H., et al., "Discovering social spammers from multiple views," *Neurocomputing*, 225, p. 49-57, 2017. Article (CrossRef Link).

[4] Moosavi, S.A., et al., "Community detection in social networks using user frequent pattern mining," *Knowledge and Information Systems*, 51(1), p. 159-186, 2017. Article (CrossRef Link).

[5] Akram, A.U., et al. "An effective experts mining technique in online discussion forums," in *Proc. of Computing, Electronic and Electrical Engineering (ICE Cube), 2016 International Conference on*. IEEE. 2016. Article (CrossRef Link).

[6] Günnemann, S., "Machine Learning Meets Databases," *Datenbank-Spektrum*, 17(1), p. 77-83, 2017. Article (CrossRef Link).

[7] Jeong, H., et al., "Detection of Zombie PCs based on email spam analysis," *KSII Transactions on Internet and Information Systems (TIIS)*, 6(5), p. 1445-1462, 2012. Article (CrossRef Link).

[8] Zhuang, X., et al., "A unified score propagation model for web spam demotion algorithm," *Information Retrieval Journal*, p. 1-28, 2017. Article (CrossRef Link).

[9] Rout, J.K., et al., "Deceptive review detection using labeled and unlabeled data," *Multimedia Tools and Applications*, 76(3), p. 3187-3211, 2017. Article (CrossRef Link).

[10] Crawford, M., et al., "Survey of review spam detection using machine learning techniques," *Journal of Big Data*, 2(1), p. 23, 2015. Article (CrossRef Link).

[11] Wang, G., et al. "Review Graph Based Online Store Review Spammer Detection," in *Proc. of 2011 IEEE 11th International Conference on Data Mining*. 2011. Article (CrossRef Link).

[12] Javanmardi, S., et al., "Fr trust: a fuzzy reputation–based model for trust management in semantic p2p grids," *International Journal of Grid and Utility Computing*, 6(1), p. 57-66, 2014. Article (CrossRef Link).

[13] Kangale, A., et al., "Mining consumer reviews to generate ratings of different product attributes while producing feature-based review-summary," *International Journal of Systems Science*, 47(13), p. 3272-3286, 2016. Article (CrossRef Link).

[14] Gani, A., et al., "A survey on indexing techniques for big data: taxonomy and performance evaluation," *Knowledge and information systems*, 46(2), p. 241-284, 2016. Article (CrossRef Link).

[15] Seneviratne, S., et al., "Spam mobile apps: Characteristics, detection, and in the wild analysis," *ACM Transactions on the Web (TWEB)*, 11(1), p. 4, 2017. Article (CrossRef Link).

[16] Page, L., et al., "The PageRank citation ranking: bringing order to the Web," 1999.

[17] Benczur, A.A., et al. "Spamrank–fully automatic link spam detection work in progress," in *Proc. of Proceedings of the first international workshop on adversarial information retrieval on the web*, 2005. Article (CrossRef Link).

[18] Li, L., et al., "Document representation and feature combination for deceptive spam review detection," *Neurocomputing*, 2017. Article (CrossRef Link).

[19] Hong, S.-S., J.-H. Kong, and M.-M. Han, "The Adaptive SPAM Mail Detection System using Clustering based on Text Mining," *KSII Transactions on Internet and Information Systems(TIIS)*, 8(6), p.2186-2196, 2014. Article (CrossRef Link).

[20] Jindal, N. and B. Liu, *Review spam detection*, in *Proceedings of the 16th international conference on World Wide Web*. 2007, ACM: Banff, Alberta, Canada. p. 1189-1190, 2007. Article (CrossRef Link).

[21] Jindal, N. and B. Liu. "Opinion spam and analysis," in *Proc. of Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM. 2008. Article (CrossRef Link).

[22] Lim, E.-P., et al. "Detecting product review spammers using rating behaviors," in *Proc. of Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010. ACM Article (CrossRef Link).

[23] Mukherjee, A., et al. "Detecting group review spam," in *Proc. of Proceedings of the 20th international conference companion on World wide web*. ACM. 2011. Article (CrossRef Link).

[24] Algur, S.P. and J.G. Biradar. "Rating consistency and review content based multiple stores review spam detection," in *Proc. of Information Processing (ICIP), 2015 International Conference on*. IEEE. 2015. Article (CrossRef Link).

[25] Lin, Y., et al., "Towards online review spam detection," in *Proc. of Proceedings of the 23rd International Conference on World Wide Web*, ACM: Seoul, Korea. p. 341-342, 2014. Article (CrossRef Link).

[26] Kumar, S., et al. "A Machine Learning Based Web Spam Filtering Approach," in *Proc. of Advanced Information Networking and Applications (AINA), 2016 IEEE 30th International Conference on*, IEEE, 2016 Article (CrossRef Link).

[27] Ye, J. and L. Akoglu. "Discovering opinion spammer groups by network footprints," in *Proc. of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2015. Article (CrossRef Link).

[28] Strötgen, J., O. Alonso, and M. Gertz. "Retro: Time-Based Exploration of Product Reviews," in *Proc. of ECIR*, Springer. 2012. Article (CrossRef Link).

[29] Chen, Y.-R. and H.-H. Chen. "Opinion spam detection in web forum: a real case study," in *Proc. of Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee*, 2015. Article (CrossRef Link).

[30] Sharma, K. and K.-I. Lin. "Review spam detector with rating consistency check," in *Proc. of Proceedings of the 51st ACM Southeast Conference*. ACM, 2013. Article (CrossRef Link).

[31] Heydari, A., M. Tavakoli, and N. Salim, "Detection of fake opinions using time series," *Expert Systems with Applications*, 58, p. 83-92, 2016. Article (CrossRef Link).

[32] Rayana, S. and L. Akoglu, "Collective Opinion Spam Detection: Bridging Review Networks and Metadata," in *Proc. of Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM: Sydney, NSW, Australia. p. 985-994, 2015. Article (CrossRef Link).

[33] Castillo, C., et al., "Know your neighbors: web spam detection using the web topology," in *Proc. of Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM: Amsterdam, The Netherlands. p. 423-430. 2007. Article (CrossRef Link).

[34] Shehnepoor, S., et al., "NetSpam: A Network-Based Spam Detection Framework for Reviews in Online Social Media*." IEEE Transactions on Information Forensics and Security*, 12(7): p. 1585-1595, 2017. Article (CrossRef Link).

[35] Xue, H. and F. Li, "A Content-Aware Trust Index for Online Review Spam Detection," in *Proc. of Data and Applications Security and Privacy XXXI: 31st Annual IFIP WG 11.3 Conference, DBSec 2017, Philadelphia, PA, USA, July 19-21, 2017, Proceedings*, G. Livraga and S. Zhu, Editors, Springer International Publishing: Cham. p. 489-508, 2017. Article (CrossRef Link).

[36] Mukherjee, A., et al. "What yelp fake review filter might be doing?" in Proc. of *ICWSM*. 2013.

[37] Heydari, A., et al., "Detection of review spam: A survey*," Expert Systems with Applications*, 42(7), p. 3634-3642, 2015.

[38] Esuli, A. and F. Sebastiani, "SentiWordNet: a high-coverage lexical resource for opinion mining*," Evaluation*, p. 1-26, 2007.

[39] Ohana, B. and B. Tierney, "Sentiment classification of reviews using SentiWordNet*,"* 2009.

[40] Hu, X., et al. "Social spammer detection with sentiment information," in *Proc. of Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE. 2014 Article (CrossRef Link).

[41] Jindal, N. and B. Liu. "Review spam detection," in *Proc. of Proceedings of the 16th international conference on World Wide Web*. ACM. 2007. Article (CrossRef Link).

[42] Krishnan, V. and R. Raj. "Web spam detection with anti-trust rank," in *AIRWeb*. 2006.

[43] Roul, R.K., et al., "Detecting spam web pages using content and link-based techniques*," Sadhana*, 41(2): p. 193-202, 2016.

[44] Abdi, H., "The Kendall rank correlation coefficient," *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks*, CA, p. 508-510, 2007.

[45] Zhang, J., M.S. Ackerman, and L. Adamic. "Expertise networks in online communities: structure and algorithms," in *Proc. of Proceedings of the 16th international conference on World Wide Web*. ACM, 2007. Article (CrossRef Link).

[46] Haveliwala, T.H., "Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search," *IEEE transactions on knowledge and data engineering*, 15(4), p. 784-796, 2003. Article (CrossRef Link).

[47] Xue, H. and F. Li. "A Content-Aware Trust Index for Online Review Spam Detection," in *Proc. of IFIP Annual Conference on Data and Applications Security and Privacy*. Springer. 2017. Article (CrossRef Link).

Mr. ABUBAKKER USMAN AKRAM received his master's degree from COMSATS Institute of Information Technology, Attock. Currently, he is pursuing his PhD degree from COMSATS institute of information and technology, WAH, Pakistan. His fields of research interest include information retrieval, scientometrics. and social network analysis.

DR. HIKMAT ULLAH KHAN received the master's degree in computer science and the Ph.D. degree in computer science from International Islamic University, Islamabad. He has been an Active Researcher for the last ten years. He is currently an Assistant Professor with the Department of Computer Science, COMSATS Institute of Information Technology, Wah Cantt, Pakistan. He has authored a number of research articles in top peer reviewed journals and international conferences. His research interests include Social web mining, Semantic Web, data science, information retrieval, and scientometrics. He is a member of the Editorial board of a number of prestigious Impact Factor Journals.

DR. SAQIB IQBAL received the M.Sc. degree in software engineering from the Queen Marry University of London and the Ph.D. degree in software engineering from the University of Huddersfield, U.K., the M.Sc. degree in computer science from Punjab University, Lahore, Pakistan and the Ph.D. degree in software engineering from the University of Huddersfield, U.K. He is currently an Assistant Professor with the Department of Software Engineering and Computer Science, College of Engineering and Information Technology, Al-Ain University of Science and Technology, Al Ain, United Arab Emirates. His research interests include software analysis and design, aspect-oriented software development, process modeling, model-based software development, requirements engineering, and design patterns in aspect-oriented programming.

DR. TASSAWAR IQBAL is presently serving as Assistant Professor at Department of Computer Science in COMSATS Institute of Information Technology, Wah, Pakistan. He completed his PhD degree from Vienna University of Technology in 2012. His current research interests include: 1) Computer Assisted Solutions for Adult Basic Education (ABE), 2) Designing learning content for ABE, 3) Impact of Computer Assisted Solutions (CAS) on ABE learners 4) Learning Styles and Multiple Intelligences in CAS, 5) Adaptive CAS, and 6) Social Networks and Societies. He has 12 publications in above mentioned areas of interests, published in reputed journals and conferences.

DR. EHSAN ULLAH MUNIR received the master's degree in computer science from the Barani Institute of Information Technology, Pakistan, in 2001, the Ph.D. degree in computer science and theory from the Harbin Institute of Technology, Harbin, China in 2008. He is currently serving as an Associate Professor and Head with the Department of Computer Science, COMSATS Institute of Information Technology, Wah Cantt, Pakistan. His research interests include heterogeneous parallel and distributed computing systems (cluster grid, cloud and peer-topeer systems), computer and wireless networks, information systems and information retrieval.

DR. MUHAMMAD SHAFI did his bachelor from Ghulam Ishaq khan institute of engineering sciences & technology and PhD from Loughborough university UK in 2005 and 2010 respectively. He has served at various universities including university of engineering & technology peshawar, university of science & technology Bannu and islamic University in medina Saudi Arabia. Currently, he is serving as associate professor at Air University Islamabad. Computer vision, machine learning, human computer interaction, mobile computing, and software engineering are his areas of research. He has also worked in software development projects for various multinational companies.