Detection of Spam on Amazon E-commerce platform

Younous Mahamat Younous[1], Yousef A.Baker El-Ebiary[2], Shadi Hilles[3]

[1]*Faculty of Computer and Information Technology, Al-Madinah International University, bougri@yahoo.fr*
[2] *Faculty of Computer and Information Technology, Al-Madinah International University, yousef.abubaker@mediu.edu.my*
[3]*Faculty of Computer and Information Technology, Al-Madinah International University, shadihilless@gmail.com*

Received 31 August 2017; accepted 11 December 2017

Abstract

The advancement of technology and the use of internet have changed many aspects of human culture over the years. Today, people look up to the internet to replace many old habits of doing things with the online shopping over platforms like Amazon and eBay becoming one of the most popular activities amongst people in modern societies. Consumers over these similar platforms take confidence in reputation and trust for comprehensive understanding of products or services when making a purchase decision. Actually well-known spam, method used to send unwanted electronic message to an unsolicited or random peoples who checks on the product or website .the major web search engine are target and easy prey use of electronic messaging systems to send an unsolicited message (spam) to make advertising on same web site .however web search engine in similar blogs social spam ,discussion groups junk fax spam mobile applications, news groups of spam .that spam is illegal but in some country they are working to legalize it somehow. The singleton is a new suggestion with hybrid classification can work forward to detect spam reviews and sentiment; E-commerce can reduce the threat of spamming on product and advertise in right way use legal channels [2]. In fact Spam had many types, so that could operate in different way.

Keywords: reviews, sentiments, singleton, E-commerce.

1. Introduction

Spamming could be viable in economic issues, because of cheap charge else could be free of charge to operate and very independent, the source of advertising, it's very common toke advantages of the technology infrastructures like IP ranges, domains.in last past years specific 2014 the volume of unsolicited mail became very high ratio the estimation is around 7 trillion worldwide .the cost of lost productivity on manufactured, company fraud and lost, their source of transmission is internet service provider. A person who creates electronic mail spam is called a spammer. Current spam on the market has well-defined objectives. Different player's web search engine services for collecting users' addresses, navigations web, searching for clients and annoying them with mass mailing.

Clearly, the internet influenced even the manner in which people express themselves and also interact with one other as well. Network of people with these virtual markets of products and services can upload their products at merchant sites e.g. amazon.com and others will post reviews on the products based on their experience with the product [2]. Some people promote or express their views on blogs and forums to promote targeted goods or services. These web content contributions from people or users are termed as user-generated content. These contents was agreed to comprise valuable information to many parties or users of the space and thus can be exploited for many reasons.

Now, not only do potential consumers search these reviews to make purchase decisions but are also used by manufacturers to identify defects in their products as well as competitive information on their potential competitors. In other words the reviews are useful to the product manufactures as they are useful to individual consumers. Example, if customer A intends to buy a product, he/she will visit this e-commerce site and browse for some existing reviews on that particular product. If he/she finds out that reviews on that product are mostly of positive opinion, then he is likely to go ahead and 3 make purchase of that product. However if these opinions are mostly negative, then there is very high probability that this consumer will choose another product. Generally, positive reviews promote products/services, bringing fame and financial gain to individuals and organizations which provided good incentives for review spamming [1]. So the statement is to:

-To identify singleton reviews with their corresponding unique reviewers.

- To Sort and label non-singleton spam reviews from dataset.

Thus the research hypothesis is by implementing ensemble of three classifiers in order to optimize efficiency in performance

2. Related work

According to all techniques involving the analysis of spamming reviews, in general at contents levels, and apply classification algorithms, like Bayesian, Support Vector Machines, and others divide spam from legal or copy right review. These approaches have been

extensively applied in spam filtering and different approach. Recently [4], designed a classifier ensemble using Naïve Bayes (NB), Support Vector Machine (SVM) and Genetic Algorithm (GA) [24]. The ensemble model shows higher percentage of classification accuracy than the base classifiers and enhances the testing time due to data dimensions reduction and significant improvement over the single classifiers.

In the past, combined logistic regression (a discriminative model) to naive Bayes (a generative model) to form an ensemble similar to the proposed study [6]. The authors observed that Naïve Bayes [4], approaches its asymptotic error without the need for a large number of training examples, and it does so very quickly. Logistic regression, on the other hand, is capable of outperforming naive Bayes, given the number of training examples is large enough [14][16]. The overall classification result was also observed to be superior to that of the base classifiers due to their respective diversity.

2.1 Naïve Bayesian Approaches

Although this approach supports only binary feature vectors, therefore is incapable of considering relevant information for the limitation in processing, they present remarkable runtime performance throughout classifying new texts. Bayesian technique can be further categorized into two groups when considering text classification, i.e. Naïve and non-Naïve Bayesian techniques. The major difference between this two is that the Naïve approach the appearance of a word within a sentence or document does not have relation with the appearance of another one within the same sentence or document. What made Naïve Bayesian approach effective in classification chore is this feature independency specification [9].

The training data is used with NB classifier for probability estimation that categorizes instances to a particular class. During training and classification stages of Naïve Bayes, small amount of storage space is needed for especially storing prior and conditional probabilities. Each message is represented as a binary vector $(x_1 \dots \dots . x_m)$, where $x_1 = 1$. If a particular token X_i of the vocabulary is present, otherwise $x_1 = 0$ [11]. The probability that a message with vector $\vec{x} = (x_1 \dots \dots \dots x_m)$ belongs in category c (= spam or lefitimate) is $(c|\vec{x}) = \frac{P(c) \times P(\vec{x}|c)}{P(\vec{x})}$ [12]. NB classifies each review in the category that maximizes the product P(c) $\times P(\vec{x}|c)$. The a priori probabilities $P(c)$ are typically estimated by dividing the number of training dataset of category c by the total number of training dataset. And the probabilities $P(\vec{x}|c)$ are calculated as follows:

$$P(\vec{x}|c) = \prod_{i=1}^{i=m} P(xi|c) = \frac{Xc+1}{Nc+|vocabulary|} \text{ where } Xc$$

is the number of occurrences of token X in reviews with label c, Nc is the total number of token occurrences in reviews labeled c and |vocabulary| is the number of unique tokens across all reviews [13].

We may write equation (1) more compactly by augmenting our feature vector such that $X_0 = 1, X = [X_0 X_1 \dots \dots \dots X_n]^T$, , and defining $w = [w_0 w_1 \dots \dots \dots \dots . . w_n]^T$. Then $s = w^T X$, the inner product between our feature *X* and weight vector *w*. Assuming our data samples are independent, we obtain the likelihood function, (*w*), such that

$$L(w) = P(Y|X,W) = \prod_{i=1}^{m} P(Y_i|X^i,w) = \prod_{i=1}^{m} \frac{(e^{-w^T X^i})(1-Y_i)}{1+e^{(-w^T X^i)}}$$

The log likelihood function is then

$$\ell(w) = \log L(w) = \sum_{i=1}^{m}(1-Y_i)\ \log\left(\frac{e^{-w^T X^i}}{1+e^{(-w^T X^i)}}\right) \quad (6)$$

$$= \sum_{i=1}^{m}(1-Y_i)\left(-w^T X^i\right) - \log\left(1+e^{-w^T X^i}\right) \quad (7)$$

We proceed by calculating the maximum likelihood estimator $\widehat{w}$, such that arg max$\{\ell(w)\}$.

Unfortunately, no closed form solution exists to calculate $\widehat{w}$. However, $\ell(w)$ is convex, so a global maximum exists and we may find it using gradient ascent. In doing so, we must move in the direction of the gradient of the function. Thus, we take the partial derivatives of $\ell(w)$ with respect to the components of *w* (note that $\frac{\partial(-w^T X^i)}{\partial w_j} = -X_j^i$)

$$= \frac{\partial \ell(w)}{\partial w_j} = \sum_{i=1}^{m} -X_j^i(1-Y^i) - X_j^i \frac{e^{-w^T X^i}}{1+e^{(-w^T X^i)}}$$

$$= \sum_{i=1}^{m} -X_j^i(1-Y^i) + X_j^i - X_j^i \frac{1}{1+e^{(-w^T X^i)}}$$

$$= \sum_{i=1}^{m} -X_j^i\left(Y^i - \frac{1}{1+e^{(-w^T X^i)}}\right)$$

$$= \sum_{i=1}^{m} -X_j^i\left(Y^i - g\left(-w^T X^i\right)\right)$$

which gives us the following gradient ascent update rule,

$$w_{new} = w_{previous} + \propto \sum_{i=1}^{m} -X_j^i\left(Y^i - g\left(-w^T X^i\right)\right)$$

2.3 General Justification Spam Review Detection Using Ensemble Classifier. On the other hand, supervised learning method to generate ensembles directly gave rise to diverse hypotheses that uses added artificially-constructed

training examples. Here the approach is easy, common meta-learner which uses any strong learner is usually used as base classifier to construct diverse committees. Experimental results employing decision-tree induction as a base learner revealed that the approach always attains higher predictive accuracy than both the base classifiers.

When considering an ensemble of classifiers, the combination of the output of a number of classifiers is only valuable if they conflict on some inputs. This study refers to the measure of disagreement as the diversity in classifier ensemble. There have been several methods proposed to measure ensemble diversity [8], usually dependent on the measure of accuracy. For regression, where the mean squared error is commonly used for determining accuracy, variance can be employed as a means for measuring diversity. So the diversity of the i^{th} classifier can be defined as

$d_i(x) = [C_i(x) - C^*(x)]^2$, Where $C_i(x)$ and $C^*(x)$ are the predictions of the i^{th} classifier and the ensemble respectively.

For this setting the generalization error, E, of the ensemble can be expressed as;

$E = \overline{E} - \overline{D}$, where $\overline{E}$ and $\overline{D}$ are the mean error and diversity of the ensemble respectively.

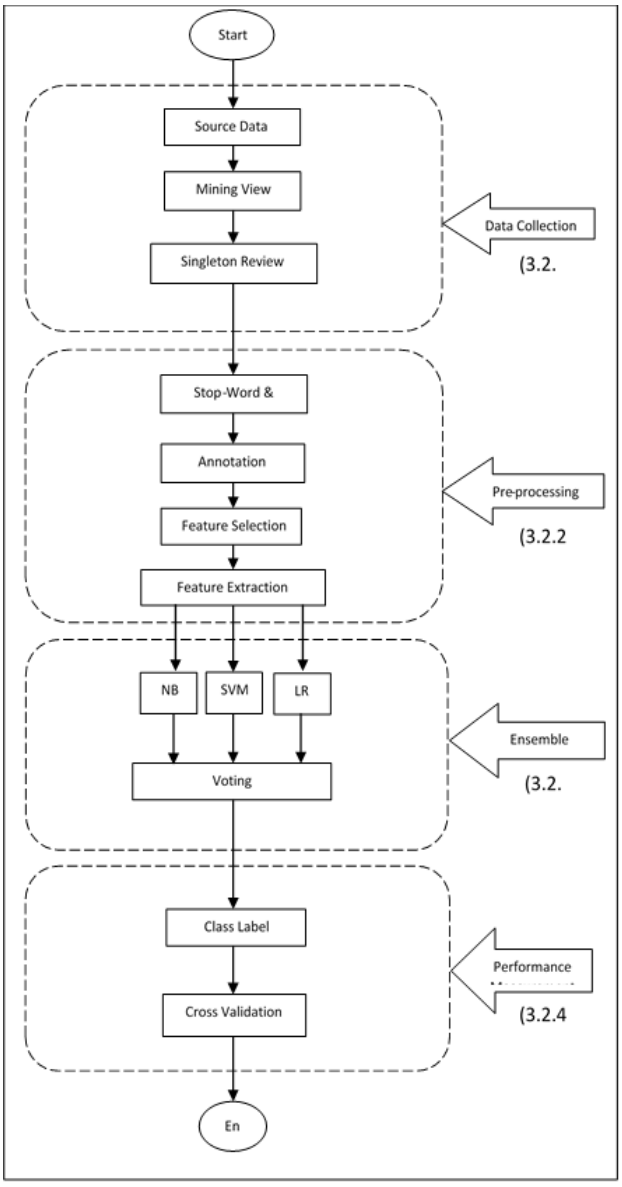
For classification problems, where the 0/1 loss function is most commonly used to measure accuracy, the diversity of the i^{th} classifier can be defined as:

$$d_i(x) = \begin{cases} 0: if\ C_i(x) = C^*(x) \\ 1: othwise \end{cases}$$

3. Methodology

Existing techniques and justifications regarding hybrid classification method for review spam detection. This part will present the model of the research, the operational framework and the methodology of designing and investigating the architecture of the hybrid classifier. It starts with the systematic framework of processes which shows the operational framework for singleton spam review detection using certain features and classification methods. After comprehensive methodology on the operational framework, the rest of the chapter deals with initiating the implementation tasks. This chapter will show the systematic framework declared for each main tasks and activities carried out in the current study. The chapter will also list all the software and hardware used in the current approach.

Figure 3.1 highlights the model of the research which includes principal tasks and activities of the current study. The input here is the dataset (reviews extracted from amazon.com) and the output is the hybrid classification system. Some preprocessing is carried out to modify the dataset in such a way it will suit the research [23]. The three classification algorithms are merged together to produce a hybrid classifier greater than their individual performance. The new ensemble classifier is then trained, using supervised machine learning and then used to filter spam reviews from the dataset. Lastly the accuracy is evaluated using standard performance measurement. Figure 3.1 will give the description and explanation of each of the main task in table for:



3.1 Performance Measurement

After successful voting to produce the final classification verdict from the ensemble classifier is been carried out the next line of action is to test its performance. This current study employs Cross Validation Technique to measure the performance of the new ensemble classifier. As explained in section (2.8.1) of previous chapter, the matrices for the evaluating the performance of the new classifier are employed. To carry out this performance measurement Spam Precision (SP), Spam Recall (SR), spam (F1) measure (F1) and accuracy were calculated. Let;

TN = number of legitimate reviews classified as legitimate (true negatives)

TP = number of spam reviews classified as spam (true positives)

FP = number of legitimate reviews classified as spam (false positives)

FN = number of spam reviews classified as legitimate (false negative)

Thus we have;

$$SP = \frac{TP}{TP + FP} \quad (3.1)$$

$$SR = \frac{TP}{TP + FN} \quad (3.2)$$

$$F1 = \frac{2 \times SP \times SR}{SP + SR} \quad (3.3)$$

$$A = \frac{TP + TN}{TP + FP + TN + FN} \quad (3.4)$$

4. Result Discussion

The research provided new insight using some particular features and hybrid classification method to detect singleton spam reviews. Most researches in spam detection focused on multiple types of spam reviews due to the variety of available features. This current research considered applying three classifiers i.e. Naïve Bayes [4], SVM and Logistic Regression parallel to solve the issue of singleton spam review. Section 4.8.1 compared the effectiveness of these individual classifiers which indicates how Hybrid classifier performed higher and more accurate than the others followed by the SVM. Therefore, the current study tried to utilize the effectiveness and vital features to carry out singleton spam review with hybrid classification using WEKA application.

Some features like; Product name, Length of the Review Title, User ID of the reviewer and Reviewer Location were totally inefficacious when compared to other features used in the research. In the other hand efficiency of Number of Helpful Feedbacks and Review Deviation from Brand Rank features was higher and has 60% of effectiveness. These features are thus introduced as suitable features for current method of classifying singleton spam reviews. The most effective features as were selected in using WEKA application respectively are: 1. rating of the Review, 2. Review Deviation from Brand Rank 3. Time Distance to Previous Review 4. Number of Helpful Feedback, 5.Length of Review Body. So in the future, other researchers would utilize these features to detect singleton type of spam.

The rest of the features which are; Rate of Product Rank, Time Distance to Next Review, Number of Feedbacks, Number of Brand Name, Review Deviation from Product Rank, Date of Publishing Review, POS Distribution, Length of Review Body, Time Distance to Previous Review and Rating of Review, had noteworthy effect on the result of classification method in this project. The effectiveness percentage of them was from 20% to 40%. Although the accuracy of a supervised method depends on various factors such as the dataset, the algorithm and the classes, a group of aforementioned features could be profitable and useful for singleton spam review detection supervised methods.

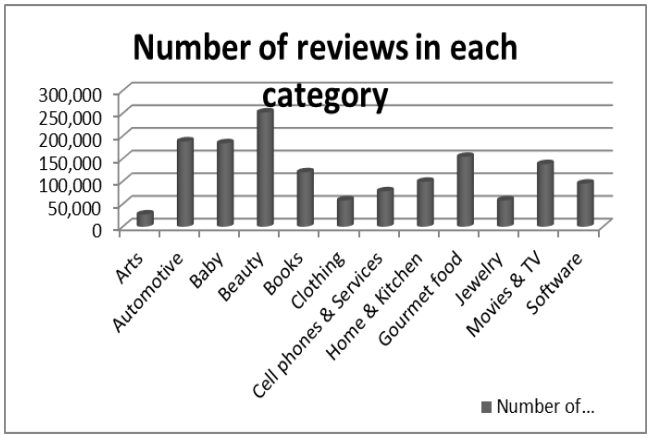
Finally, among three proposed features, just one of them (Time Distance to Previous Review) was selected as a member of the best combination of features for classification method for singleton spam review detection. The two proposed features: Time Distance to Next Review and Num. of Brand Name had some effect in this current approach. However it was expected that all of the reasons for obtaining results for the features might be the limitations in the dataset. A few types of spam behaviors were observed in the dataset used for current research. It obviously changes the effectiveness of utilized features. Therefore, in order to access the effectiveness of proposed features a large gold standard dataset must be provided. Then the exact role of features the main tools phishers use for their malicious purposes. In recent years, spam adverting the chance to make a profit from the shares of various listed companies has become more prevalent. This type of spam is therefore most probably initiated by those who gamble on the stock markets in an attempt to influence stock prices in their favor. The situation in Russia is slightly different. The spammers' services are often employed by representatives of small businesses trading in such things as electronics, spare parts, cars, legal services, tourism, medicine, etc. in order to increase company turnover.

4.1 Dataset contents

No	Item	No	Item
1	Arts product	14	Jewelry
2	Automotive	15	Office products
3	Baby	16	Pet supply
4	Books	17	Music
5	Beauty	18	Musical instruments
6	Clothing	19	Sports and Outdoor products
7	Cell phones and accessories	20	Shoe
8	Home and Kitchen products	21	Software
9	Industrial and scientific products	22	Watch
10	Electronics product	23	Tools and hardware
11	Gourmet food	24	Toys and games

12	Movie and TV	25	Video game
13	Health products		

4.2 Number of reviews in each category



4.3 Mining View

Preparation tasks were made when collecting the dataset to customize the data in order to suit the research. Collected the first version of the dataset from amazon.com and it was in the form of TXT file. Source data section also clarified that the dataset contains many reviews in multifarious categories. There is also difficulty in manually selecting reviews from Movies & TV category or deletion of reviews from other categories.

To prepare the corpus in a practicable and suppler manner, DOM Parser was used to parse the XML version of the dataset in Java. Before the dataset XML file is being processed in java, there is need to validate its content. Therefore, to validate the file, some manual and computational processes were applied on the XML file.

4.4 Dataset Annotation

In this phase, the processed dataset will be analyzed and labeled to be used as training data for the classification approach. Manual labeling of training data is a gruesome task which takes time thus complicating supervised methods of spam review detection. These prompt the idea that annotating the data to be used as training data by ordinary individuals will not be adequate. This study implemented a programmed method of carrying out the annotation towards preparing the training data [4].

In the GATE application, spam reviews are considered to have the value of 1 and non-spam reviews are considered to have the value of 0. Is a screenshot illustrating how annotation is being carried out in the gate application. To extract POS (Part of Speech) distribution feature using GATE application, the ANNIE Plugin was firstly added and then followed by three processing resources namely; ANNIE English Tokenizer, ANNIE Sentence Splitter and ANNIE POS Tagger were added. After those processes the ANNIE POS Tagger application was r. This results to the successful extraction of the POS distribution from the dataset.

4.5 Founding result discussion of classifier

The research provided new insight using some particular features and hybrid classification method to detect singleton spam reviews. Most researches in spam detection focused on multiple types of spam reviews due to the variety of available features. This current research considered applying three classifiers i.e. Naïve Bayes, SVM and Logistic Regression parallel to solve the issue of singleton spam review. Compared the effectiveness of these individual classifiers which indicates how Hybrid classifier performed higher and more accurate than the others followed by the SVM, in the cloud computing for virtual machines [23] is described Moving to the cloud from a virtual environment. Therefore, the current study tried to utilize the effectiveness and vital features to carry out singleton spam review with hybrid classification using WEKA application. The corresponding result was represented

5. Results from Classifier Selection for Ensemble

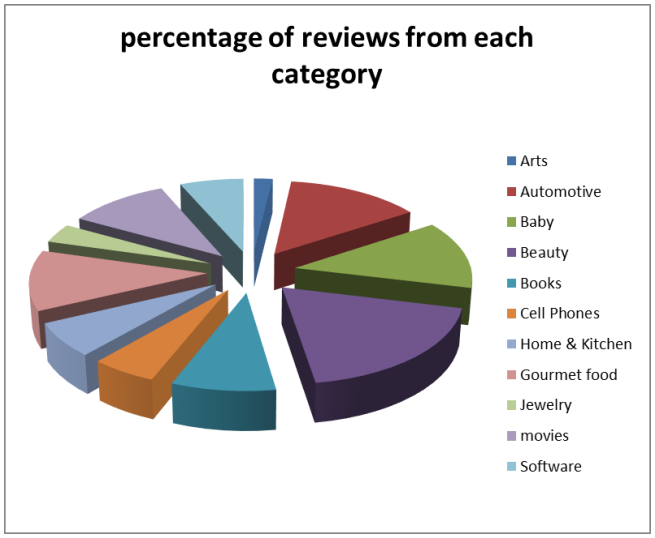
As described in the previous chapter, an ensemble would only be more successful than an individual classifier if the individual classifiers disagree with each other. Such independence among classifiers is known as the diversity within an ensemble. Bagged models of various classifiers was used for the bagging phase but only Data Preparation and Pre-processing Results and Analysis. This section presents the result from data preparation and pre-processing in the activities of the project. The first step in building a classifier is to transform the reviews into a form that is acceptable or recognized by the classifier algorithms. These processes include:

.Tokenize the review text and establish an initial list of terms. Eliminate stop-words using a pre-defined stop list.

.Perform stemming with variant of Porter algorithm. This approach or results declared in this phase are called the initial findings because they are applied before any phase of the study.

5.1 Result of Data Preparation

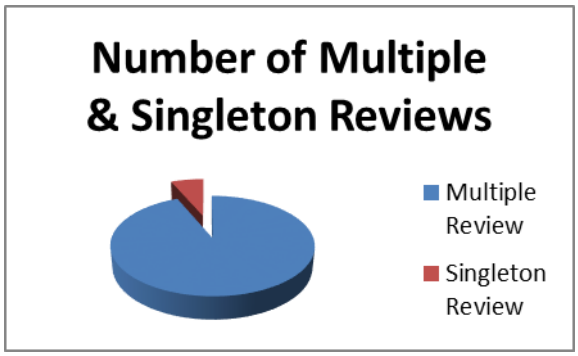
The dataset applied for this study contains consumer reviews on Movies & TV category of products collected from amazon.com [22]. The dataset contains 25 categories of products and each category containing many numbers of related products with numerous reviews posted for each products. Figure 4.1 is a pie chart that illustrates the number in percentage of reviews from each category (only 10 out of 25 are shown). As explained in the previous chapter, 24 categories out of the 25 categories were deleted to come up with reviews on IMDB movies that were used for the current research.



Initially, there were 148,408 reviews in the movie category and 25 categories in the original dataset. To prepare the dataset in a practicable and suppler manner, DOM Parser was used to parse the XML version of the dataset in Java. Before the dataset XML file is being processed in java, there is need to validate its content. Therefore, to validate the file, some manual and computational processes were applied on the XML file. After successful computation the DOM Parser is used to parse the file in Java. Also additional decision in java was used to help discard reviews from the 25 categories, leaving only movie category [22].

5.2 Number of Multiple and Singleton Reviews

Type of Review Number	Total 148, 408	Multiple Review 146,342	Singleton Review 2,066



Initially the dataset was processed to have 1089 healthy reviews and 968 spam reviews. Further adjustments had been taken on this dataset later to produce exactly 1000 spam reviews and another 1000 healthy reviews.

6. CONCLUSION

It's clear that spam detection has significant key role in business domains. E-commerce and opinion sharing websites changed the way people purchase products and convey their opinions on them. In such situations, customers share their opinion on products by posting reviews in aforementioned websites. Certainly a set of reviews on a product is valuable for potential customers to purchase a product very close to their needs and business holders to update their products, analyze their weakness and control the market. So the accuracy of these sources of information which could be realized using review spam detection techniques is profoundly effective on business.

References

[1]. JINDAL, N. & LIU, B. Opinion spam and analysis.Proceedings of the 2013 International Conference on Web Search and Data Mining, 2008. ACM, 219-230.

[2]. MUDAMBI, S. M. & SCHUFF, D. 2010. What makes a helpful online review? A study of customer reviews on Amazon. com. Management Information Systems Quarterly, 34, 3 BARIGOU, F., BARIGOU, N. & ATMANI, B. 2012. Combining Classifiers for Spam Detection. Networked Digital Technologies. Springer.

[3]. BARIGOU, F., BARIGOU, N. & ATMANI, B. Spam Detection System Combining Cellular Automata and Naïve Bayes Classifier. ICWIT, 2014. Citeseer, 250-260.

[4]. [4]. BAUTIN, M., VIJAYARENU, L. & SKIENA, S. International Sentiment Analysis for News and Blogs. ICWSM, 2012.

[5]. [5].BENCZUR, A. A., CSALOGANY, K., SARLOS, T. & UHER, M. Spam Rank–Fully Automatic Link Spam Detection Work in progress. Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, 2011.

[6]. BRÜCHER, H., KNOLMAYER, G. & MITTERMAYER, M.-A. 2012. Document classification methods for organizing explicit knowledge. Research Group Information Engineering, Institute of Information Systems, University of Bern, Engehaldenstrasse, 8.

[7]. CASTILLO, C. & DAVISON, B. D. 2013. Foundations and Trends® in Information Retrieval. Foundations and Trends® in Information Retrieval, 4, 377-486.

[8]. DAVE, K., LAWRENCE, S. & PENNOCK, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. Proceedings of the 12th international conference on World Wide Web, 2003. ACM, 519-528.

[9]. DIETTERICH, T. G. 2013. Ensemble methods in machine learning. Multiple classifier systems. Springer.

[10]. HOSE, A., IPEIROTIS, P. G. & SUNDARARAJAN, A. Opinion mining using econometrics: A case study on reputation systems. Annual Meeting-Association for Computational Linguistics, 2013. Citeseer, 416.

[11]. GRIES, S. T. & BEREZ, A. L. Linguistic annotation in/for corpus linguistics. Handbook of Linguistic Annotation. Berlin, New York: Springer. Abgerufen von http://www. Linguistics. Ucsb. edu/faculty/stgries/research/InProgr_STG_ALB_LingAnnotCorpLing_HbOfLingAnnot. Pdf.

[12]. GUPTA, M. & AGGARWAL, N. Classification techniques analysis. NCCI2010-National Conference on Computational Instrumentation, CSIO Chandigarh, India, 2010. 19-20.

[13]. HANCOCK, J. T., CURRY, L. E., GOORHA, S. & WOODWORTH, M. 2007. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. Discourse Processes, 45, 1-23.

[14]. HUANG, G.-B., WANG, D. H. & LAN, Y. 2011. Extreme learning machines: a survey. International Journal of Machine Learning and Cybernetics, 2, 107-122.
[15]. JINDAL, N. & LIU, B. Mining comparative sentences and relations. AAAI, 2014. 1331-1336.
[16]. JINDAL, N. & LIU, B. Analyzing and detecting review spam. Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, 2013a. IEEE, 547-552.
[17].].JINDAL, N. & LIU, B. Review spam detection. Proceedings of the 16th international conference on World Wide Web, 2013b. ACM, 1189-1190.
[18]. JINDAL, N. & LIU, B. Opinion spam and analysis. Proceedings of the 2012 International Conference on Web Search and Data Mining, 2014. ACM, 219-230.
[19]. JORDAN, A. 2012. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems, 14, 841.
[20]. .MCAULEY, J. & LESKOVEC, J. Hidden factors and hidden topics: understanding rating dimensions with review text. Proceedings of the 7th ACM conference on Recommender systems, 2013. ACM, 165-172.
[21]. RAINA, R., SHEN, Y., MCCALLUM, A. & NG, A. Y. Classification with hybrid generative/discriminative models. Advances in neural information processing systems, 2003. None.
[22]. MUTHUKUMARASAMY, G. 2014. Spam Review Detection using a Hybrid Classification Method. International Journal of Advances in Engineering Sciences, 4, 22
[23]. Hassan, S. S. M., & Hilles, S. M. (2014). Enhancing Security Concerns in Cloud Computing Virtual Machines:(Case Study on Central Bank of Sudan).