International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 08 | Aug 2022 www.irjet.net p-ISSN: 2395-0072

Fake Reviews Detection using Supervised Machine Learning

B. Jyoshna[1], G. Somasekhar[2], Dr. M. Saravanamuthu[3]

[1]Student, Department of Computer Applications, Madanapalle institute of technology and science, India
[2]Student, Department of Computer Applications, Madanapalle institute of technology and science, India
[3]Asst. Professor, Department of Computer Applications, Madanapalle institute of technology and science, India

---

ABSTRACT - *With the non-stop evolve of E-commerce systems, on-line opinions are mostly regarded as a critical component for building and retaining a proper reputation. Moreover, they have an tremendous function in the selection making procedure for give up users. Usually, a high-quality evaluate for a goal object attracts greater clients and lead to excessive extend in sales. Nowadays, misleading or faux opinions are intentionally written to construct digital popularity and attracting conceivable customers. Thus, figuring out pretend evaluations is a vivid and ongoing lookup area. Identifying faux opinions relies upon now not solely on the key aspects of the evaluations however additionally on the behaviors of the reviewers. This paper proposes a computing device mastering strategy to perceive pretend reviews. In addition to the elements extraction technique of the reviews, this paper applies numerous facets engineering to extract a number of behaviors of the reviewers. The paper compares the overall performance of countless experiments accomplished on a actual Yelp dataset of eating places critiques with and barring facets extracted from customers behaviors. In each cases, we examine the overall performance of a number of classifiers; KNN, Naive Bayes (NB), SVM, Logistic Regression and Random forest. Also, extraordinary language fashions of n-gram in precise bi-gram and tri-gram are taken into issues for the duration of the evaluations. The effects expose that KNN(K=7) outperforms the relaxation of classifiers in phrases of f-score reaching exceptional f-score 82.40%. The effects exhibit that the f-score has improved by using 3.80% when taking the extracted reviewers' behavioral points into consideration.*

Keywords - Fake reviews detection; data mining; supervised machine learning

1.INTRODUCTION

Nowadays, when clients choose to draw a choice about offerings or products, critiques end up the predominant supply of their information. For example, when clients take the initiation to e book a hotel, they examine the opinions on the opinions of different clients on the motel services. Depending on the comments of the reviews, they determine to e book room or not. If they got here to high quality remarks from the reviews, they in all likelihood proceed to e book the room. Thus, historic opinions grew to be very credible sources of records to most humans in a number of on-line services. Since, critiques are viewed varieties of sharing actual remarks about fantastic or poor services, any strive to manipulate these critiques through writing deceptive or inauthentic content material is regarded as misleading motion and such opinions are labeled as faux [1].

Discover and extract beneficial Such case leads us to suppose what if now not all the written critiques are truthful or credible. What if some of these critiques are fake. Thus, detecting pretend overview has end up and nevertheless in the nation of lively and required lookup vicinity [2].

Machine mastering methods can grant a large contribution to realize pretend evaluations of net contents. Generally, internet mining methods [3] records the usage of various computer getting to know algorithms. One of the internet mining duties is content material mining. A normal instance of content material mining is opinion mining [4] which is involved of discovering the sentiment of textual content (positive or negative) by using desktop gaining knowledge of the place a classifier is educated to analyze the facets of the evaluations collectively with the sentiments. Usually, pretend critiques detection relies upon no longer solely on the class of critiques however additionally on positive elements that are now not immediately related to the content. Building elements of evaluations usually entails textual content and herbal language processing NLP. However, pretend opinions may additionally require constructing different elements linked to the reviewer himself like for instance evaluation time/date or his writing styles. Thus, the profitable pretend critiques detection lies on the development of significant aspects extraction of the reviewers.

To this end, this paper applies several machine learning classifiers to identify fake reviews based on the content of the reviews as well as several extracted features from the reviewers. We apply the classifiers on real corpus of reviews taken from Yelp [5]. Besides the normal natural language processing on the corpus to extract and feed the features of the evaluations to the classifiers, the paper additionally applies countless points engineering on the corpus to extract a number of behaviors of the reviewers. The paper compares the influence of extracted elements of the reviewers if they are taken into consideration inside the classifiers. The papers examine the effects in the absence and the presence of the extracted facets in two special language fashions particularly TF-IDF with bi-grams and TF-IDF with tri-grams. The outcomes shows that the engineered elements amplify the overall performance of pretend evaluations detection process. The relaxation of this paper is prepared as follows: Section II Summarizes the kingdom of artwork in detecting pretend reviews. Section III introduces a historical past about the laptop gaining knowledge of techniques. Section IV provides the small print of the proposed approach. Conclusions and future work are brought in SectionVI.

2.PROPOSEDAPPROACH

This area explains the small print of the proposed strategy proven in discern 1. The proposed strategy consists of three fundamental phases in order to get the first-class mannequin that will be used for faux critiques detection. These phases are defined in the following:

A. Data Preprocessing

The first step in the proposed method is information preprocessing [26]; one of the integral steps in desktop gaining knowledge of approaches. Data preprocessing is a vital undertaking as the world facts is in no way suitable to be used. A sequence of preprocessing steps have been used in this work to put together the uncooked facts of the Yelp dataset for computational activities. This can be summarized as follows:

1) Tokenization: Tokenization is one of the most frequent herbal language processing techniques. It is a primary step earlier than making use of any different preprocessing techniques. The textual content is divided into man or woman phrases referred to as tokens. For example, if we have a sentence ("wearing helmets is a ought to for pedal cyclists"), tokenization will divide it into the following tokens ("wearing" , "helmets" , "is" , "a", "must", "for" , "pedal" , "cyclists") [27].

2) Stop Words Cleaning: Stop phrases [28] are the phrases which are used the most but they preserve no value. Common examples of the give up phrases are (an, a, the, this). In this paper, all statistics are cleaned from cease phrases earlier than going ahead in the pretend opinions detection process.

3) Lemmatization: Lemmatization approach is used to con- vert the plural structure to a singular one. It is aiming to cast off inflectional endings solely and to return the base or dictionary structure of the word. For example: changing the phrase ("plays") to ("play") [29].

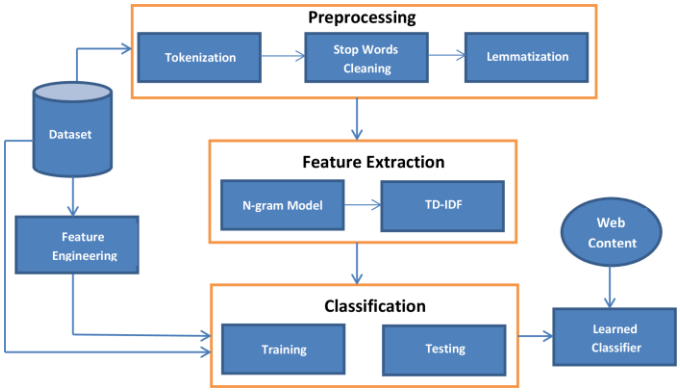

Fig. 1. The Proposed Framework.

A. B Feature Extraction

Feature extraction is a step which ambitions to amplify the overall performance both for a sample cognizance or laptop studying system. Feature extraction represents a discount segment of the information to its essential elements which yields in feeding computer and deep getting to know fashions with greater precious data. It is commonly a manner of doing away with the unneeded attributes from facts that may additionally absolutely decrease the accuracy of the mannequin.

Several procedures have been developed in the literature to extract points for faux critiques detection. Textual points is one famous method [31]. It consists of sentiment classification[32] which relies upon on getting the percentage of fantastic and terrible phrases in the review; e.g. "good", "weak". Also, the Cosine similarity is considered. The Cosine similarity is the cosine of the attitude between two n-dimensional vectors in an n-dimensional house and the dot product of the two vectors divided by means of the product of the two vectors' lengths (or magnitudes)[33]. TF-IDF is some other textual characteristic technique that receives the frequency of

each genuine and false (TF) and the inverse record (IDF). Each phrase has a respective TF and IDF rating and the product of the TF and IDF ratings of a time is known as the TF-IDF weight of that time period [34]. A confusion matrix is used to classify the opinions into 4 results; True Negative (TN): Real activities are categorised as actual events, True Positive (TP): Fake occasions are categorised as fake, False Positive (FP): Real occasions are categorized as faux events, and False Negative (FN): Fake activities are categorised as real.

Second there are person private profile and behavioural features. These aspects are the two methods used to pick out spammers Whether through the usage of time-stamp of user's remark is established and special than different ordinary customers or if the person posts a redundant evaluate and has no relation to area of target. In this paper, We observe TF-IDF to extract the facets of the contents in two languages models; commonly bi-gram and tri-gram. In each language models, we follow additionally the prolonged dataset after extracting the points representing the customers behaviours.

A. Feature Engineering

Fake reviews are known to have other descriptive features [35] related to behaviors of the reviewers during writing their reviews. In this paper, we consider some of these feature and their impact on the performance of the fake reviews detection process. We consider caps-count, punct-count, and emojis behavioral features. caps-count represents the total capital character a reviewer use when writing the review, punct-count represents the total number of punctuation that found in each review, and emojis counts the total number of emojis in each review. Also, we have used statistical analysis on reviewers behaviors by applying "groupby" function, that gets the number of fake or real reviews by each reviewer that are written on a certain date and on each hotel. All these features are taken into consideration to see the effect of the users behaviors on the performance of the classifiers.

3.EXPERIMENTAL RESULTS

We evaluated our proposed device on Yelp dataset [5]. This dataset consists of 5853 opinions of 201 inns in Chicago written with the aid of 38, sixty-three reviewers. The evaluations are categorized into 4, 709 assessments labeled as actual and 1, a hundred and forty-four opinions labeled as fake. Yelp has categorized the opinions into actual and fake. Each occasion of the evaluate in the dataset consists of the evaluation date, evaluation ID, reviewer ID, product ID, evaluation label and megastar rating. The statistic of dataset is summarized in Table I. The most assessment size in the information includes 875 words, the minimal evaluate size carries four words, the common size of all the critiques is 439.5 word, the whole quantity of tokens of the facts is 103052 words, and the variety of special phrases is 102739 word.

TABLE -1. Summary of the dataset

Total number of reviews	5853
review Number of fake reviews	1144 review
Number of real reviews	4709 review
Number of distinct words	102739-word
Total number of tokens	103052
token	
The Maximum review length	875 word
The Minimum review length	4 word
The Average review length	439.5 word

TABLE - 2. Accuracy of bi-gram and tri-gram in the absence of

Classification Algorithm	Extracted Features Behaviors Bigram Accuracy	Accuracy % Trigram	Average Accurac
% Logistic Regression	**87.87%**	**87.87%**	**87.87%**
Naive bayes	86.76%	87.30%	87.03%
KNN (K=7)	86.34%	**87.87%**	87.82%
SVM	87.82%	87.82%	87.82%
Random Forest	87.82%	87.82%	87.82%

In addition to the dataset and its statistics, we extracted other features representing the behaviors of reviewers during writing their reviews. These features include *caps-count* which represents the total capital character a reviewer use when writing the review, *punct-count* which represents the total number of punctuations that found in each review, and *emojis* which counts the total number of emojis in each review. We will take all these features into consideration to see the effect of the users behaviors on the performance of the classifiers.

In this part, we present the results for several experiments and their evaluation using five different machine learning classifiers. We first apply TF-IDF to extract the features of the contents in two languages models; mainly bi-gram and tri- gram. In both language models, we apply also the extended dataset after extracting the features representing the users behaviors mentioned in the last section. Since the dataset is unbalanced in terms of positive and negative labels, we take into consideration the precision and the recall, and hence and hence f1-score is considered as a

performance measure inaddition to accuracy. 70% of the dataset is used for training while 30% is used for testing. The classifiers are first evaluated in the absence of extracted features behaviors of users and then in the presence of the extracted behaviors. In each case, we compare the performance of classifiers in Bi-gram and Tri-gram language models.

Table II Summarizes the results of accuracy in the absence of extracted features behaviors of users in the two language models. The average accuracy for each classifier of the two language models is shown. It is found that the logistic re- gression classifier gives the highest accuracy of 87.87% in Bi-gram model. SVM and Random forest classifiers have relatively close accuracy to logistic regression. In Tri-gram model, KNN and Logistic regression are the best with accuracyof 87.87%. SVM and Random forest have relatively closeaccuracy with score of 87.82%. In order to evaluate the overall performance, we take into consideration the average accuracy of each classifier in both language models. It is found that the highest average accuracy is achieved in logistic regression with 87.87%. The summary of the results are shown in Fig. 2.

On the other hand, Table III summarizes the accuracy of the classifiers in the presence of the extracted features behaviorsof the users in the two language models. The results reveal that the classifiers that give the highest accuracy in Bi-gramis SVM with score of 86.9%. Logistic regression and Random Forest have relativity close accuracy with score of 86.89% and 86.85%, respectively. While in Tri-gram model, both SVM, and logistic regression give the best accuracy with score of 86.9%. The Random Forest gives a close score of 86.8%. The summary of the results is illustrated in Fig. 3. Also, it isfound that the highest average accuracy is obtained with SVM classifier with score of 86.9%.

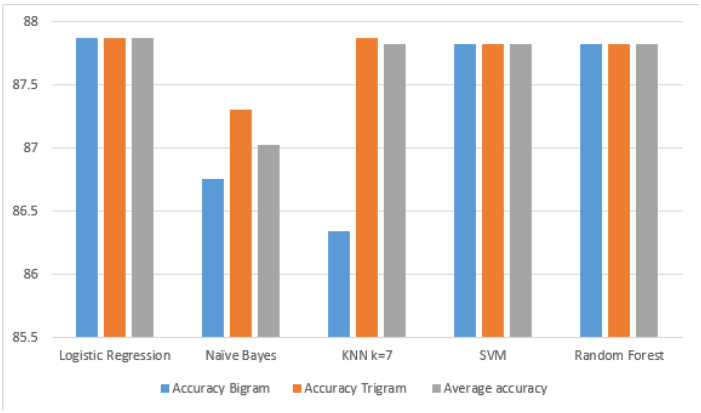

Fig -2. Accuracy, and Average Accuracy in Absence of Extracted Behavioral Features.

Additionally, precision, Recall and f1-score are taken into consideration as evaluation metrics. Actually, they are key indicators when the data is unbalanced similar to the previous, table 4 represents the recall, precision, and hence f-score in the absence of the extracted features behaviors of the user in the two language models. For the trade off between recall and precision, f1-score is taken into account as the evaluation criterion of each classifier. In bi-gram, KNN(k=7) outperforms all other classifiers with f1-score Value of 82.40%. Whereas, in Tri-gram, both logistic regression and KNN(K-7) outperform other classifiers with f1-score value of 82.20%. To evaluate the overall performance of the classifiers in both language models,

TABLE - 3. Accuracy of bi-gram and tri-gram in the presence of extracted features behaviors

Classification Algorithm	Accuracy% Bigram	Accuracy% Trigram	Average Accuracy
Logistic Regression	86.89%	**86.9%**	86.89%
Naive bayes	85.82%	86.34%	86.08%
KNN (K=7)	86.56%	85.9%	86.23%
SVM	**86.9%**	**86.9%**	**86.9%**
Random Forest	86.85%	86.8%	86.82%

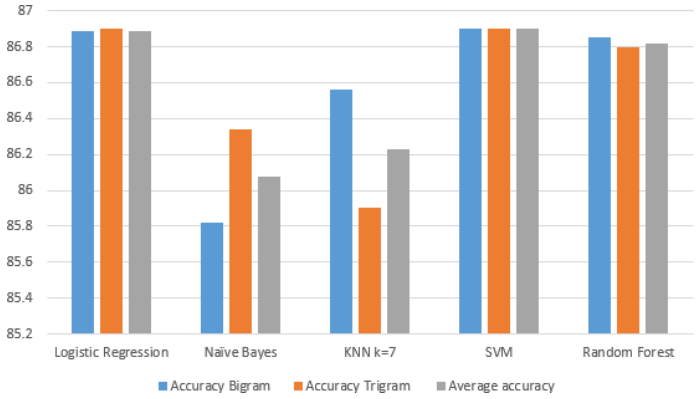

Fig -3. The Accuracy, and the Average Accuracy after Applying Feature Engineering

The average f1-score is calculated. It is found that, KNN outperforms the overall classifiers with average f1-score of 82.30%. Fig. 4 depicts the the overall performance of all classifiers.

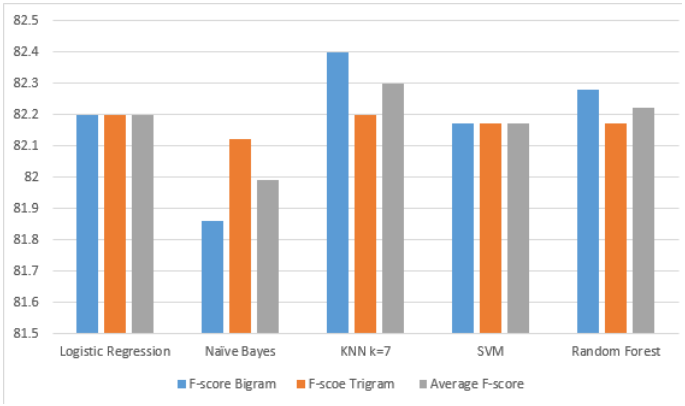

Fig -4. f-score, and Average f-score in Absence of Extracted Behavioral Features.

TABLE IV. RECALL, PRECISION, AND F1-SCORE IN ABSENCE OF EXTRACTED BEHAVIORAL FEATURES

	Bi-gram			Tri-gram			Avg F-score
	Recall	Precision	F-score	Recall	Precision	F-score	
Logistic Regression	**87.87%**	77.22%	82.20%	**87.87%**	77.20%	**82.20%**	82.20%
Naive Bayes	86.79%	78.23%	81.86%	87.30%	**78.97%**	82.12%	81.99%
KNN(K=7)	86.34%	80.20%	**82.40%**	**87.87%**	77.22%	**82.20%**	**82.30%**
SVM	87.82%	77.21%	82.17%	87.82%	77.21%	82.17%	82.17%
Random Forest	87.82%	**81.29%**	82.28%	87.82%	77.21%	82.17%	82.22%

Similarly, Table V summarizes the recall, precision, and f1- score in the presence of the extracted features behaviors of the users in the two language models. It is found that, the highest f1-score value is achieved by Logistic regression with f1-scorevalue of 82% in case of Bi-gram. While the highest f1-score value in Tri-gram is achieved in KNN with f1-score value of 86.20%. Fig. 5 illustrates the performance of all classifiers. The KNN classifier outperforms all classifiers in terms of the overall average f1-score with value of 83.73%.

The results reveal that KNN(K=7) outperforms the rest of classifiers in terms of f-score with the best achieving f- score 82.40%. The result is raised by 3.80% when taking the extracted features into consideration giving best f-score value of 86.20%

TABLE V. RECALL, PRECISION, AND F1-SCORE IN PRESENCE OF EXTRACTED BEHAVIORAL FEATURES

	Bi-gram			Tri-gram			Avg F-score
	Recall	Precision	F-score	Recall	Precision	F-score	
Logistic Regression	**86.90%**	75.53%	**82%**	86.90%	75.53%	80.82%	81.41%
Naive Bayes	85.82%	76%	80.38%	86.34%	76.59%	80.64%	80.51%
KNN(K=7)	86.56%	**80%**	81.26%	85.30%	**78.50%**	**86.20%**	**83.73%**
SVM	**86.90%**	75.50%	80.82%	84.90%	75.53%	81.82%	81.32%
Random Forest	86.85%	75.50%	80.79%	**87.90%**	74.53%	81.90%	81.34%

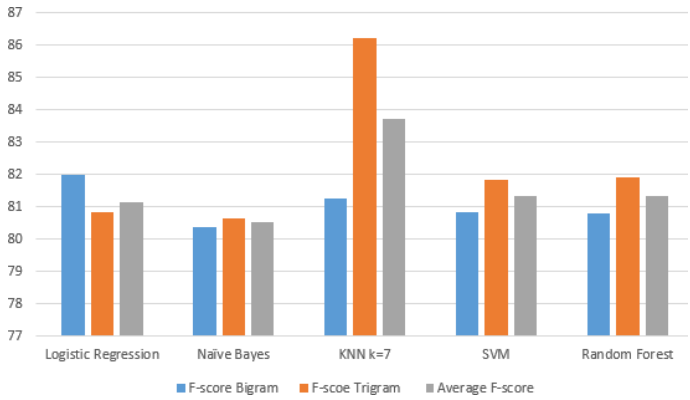

Fig -5. f-score, and Average f-score in Presence of Extracted BehavioralFeatures.

It is obvious that reviews play a crucial role in people's decision. Thus, fake reviews detection is a vivid and ongoing research area. In this paper, a machine learning fake reviews detection approach is presented. In the proposed approach, both the features of the reviews and the behavioral features of the reviewers are considered. The Yelp dataset is used to evaluate the proposed approach. Different classifiers are implemented in the developed approach. The Bi-gram and Tri- gram language models are used and compared in the developed approach. The results reveal that KNN(with K=7) classifier outperforms the rest of classifiers in the fake reviews detection process. Also, the results show that considering the behavioral features of the reviewers increase the f-score by 3.80%. Not all reviewers behavioral features have been taken into consideration in the current work. Future work may consider including other behavioral features such as features that depend on the frequent times the reviewers do the reviews, the time reviewers take to complete reviews, and how frequent they are submitting positive or negative reviews. It is highly expected that considering more behavioral features will enhance the performance of the presented fake reviews detection approach.

4.CONCLUSION

In this paper, we showed the importance of reviews and how they affect almost everything related to web based data.

ACKNOWLEDGMENTS

The authors would like to thank the Deanship of Scientific Research in Prince Sattam Bin Abdelaziz

University, KSA for his support during the stages of this research.

REFERENCS

[1] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *Information Processing & Management*, vol. 56, no. 4, pp. 1234 – 1244, 2019.

[2] S. Tadelis, "The economics of reputation and feedback systems in e- commerce marketplaces," *IEEE Internet Computing*, vol. 20, no. 1, pp. 12–19, 2016.

[3] M. J. H. Mughal, "Data mining: Web data mining techniques, tools and algorithms: An overview," *Information Retrieval*, vol. 9, no. 6, 2018.

[4] C. C. Aggarwal, "Opinion mining and sentiment analysis," in *Machine Learning for Text*. Springer, 2018, pp. 413–434.

[5] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What yelp fake review filter might be doing?" in *Seventh international AAAI conference on weblogs and social media*, 2013.

[6] N. Jindal and B. Liu, "Review spam detection," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07, 2007.

[7] E. Elmurngi and A. Gherbi, *Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques*. IARIA/DATA ANA- LYTICS, 2017.

[8] V. Singh, R. Piryani, A. Uddin, and P. Waila, "Sentiment analysis of movie reviews and blog posts," in *Advance Computing Conference (IACC)*, 2013, pp. 893–898.

[9] A. Molla, Y. Biadgie, and K.-A. Sohn, "Detecting Negative Deceptive Opinion from Tweets." in *International Conference on Mobile and Wireless Technology*. Singapore: Springer, 2017.

[10] S. Shojaee *et al.*, "Detecting deceptive reviews using lexical and syntactic features." 2013.

[11] Y. Ren and D. Ji, "Neural networks for deceptive opinion spam detection: An empirical study," *Information Sciences*, vol. 385, pp. 213–224, 2017.

[12] H. Li *et al.*, "Spotting fake reviews via collective positive-unlabeled learning." 2014.

[13] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ser. WSDM '08, 2008, pp. 219–230.

[14] 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," *Journal of Management Information Systems*, vol. 33, no. 2, pp. 456–481, 2016.

[15] E. D. Wahyuni and A. Djunaidy, "Fake review detection from a product review using modified method of iterative computation framework." 2016.

[16] D. Michie, D. J. Spiegelhalter, C. Taylor *et al.*, "Machine learning,"

Neural and Statistical Classification, vol. 13, 1994.

[17] T. O. Ayodele, "Types of machine learning algorithms," in *New ad- vances in machine learning*. InTech, 2010.

[18] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[19] T. Joachims, "Text categorization with support vector machines: Learn- ing with many relevant features." 1998.

[20] T. R. Patil and S. S. Sherekar, "Performance analysis of naive bayes and j48 classification algorithm for data classification," pp. 256–261, 2013.

[21] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048,2007.

[22] N. Suguna and K. Thanushkodi, "An improved k-nearest neighbor clas- sification using genetic algorithm," *International Journal of Computer Science Issues*, vol. 7, no. 2, pp. 18–21, 2010.

[23] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote sensing of environment*, vol. 61, no. 3, pp. 399–409, 1997.

[24] A. Liaw, M. Wiener *et al.*, "Classification and regression by random- forest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[25] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic regression*. Springer, 2002.

[26] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, no. 1, pp. 51–89, 2003.

[27] J. J. Webster and C. Kit, "Tokenization as the initial phase in nlp," in *Proceedings of the 14th conference on Computational linguistics- Volume 4*. Association for Computational Linguistics, 1992, pp. 1106–

1110.

[28] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization," in *Neural Networks, 2003. Proceedings of the International Joint Conference on*, vol. 3. IEEE, 2003, pp. 1661–1666.

[29] J. Plisson, N. Lavrac, D. Mladenic´ *et al.*, "A rule based approach to word lemmatization," 2004.

[30] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Transactions on Pattern Analysis & Machine Intel- ligence*, no. 4, pp. 388–400, 1993.

[31] N. Jindal and B. Liu, "Opinion spam and analysis." in *Proceedings of the 2008 international conference on web search and data mining*. ACM, 2008.

[32] M. Hu and B. Liu, "Mining and summarizing customer reviews." 2004.

[33] R. Mihalcea, C. Corley, C. Strapparava *et al.*, "Corpus-based and knowledge-based measures of text semantic similarity," in *AAAI*, vol. 6, 2006, pp. 775–780.

[34] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machinelearning*, vol. 242, 2003, pp. 133–142.

[35] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting burstiness in reviews for review spammer detection," in *Seventh international AAAI conference on weblogs and social media*, 2013.