Comparison of Machine Learning Algorithms for Spam Detection

Azeema Sadia [1,*], Fatima Bashir [1], Reema Qaiser Khan [1], Amna Bashir [2], and Ammarah Khalid [3]

[1] Bahria University, Dept. of Computer Science, Karachi, Pakistan
[2] Sir Syed University of Engineering & Technology, Dept. of Software Engineering, Karachi, Pakistan
[3] Bahria University, Dept. of Software Engineering, Karachi, Pakistan
*Correspondence: azeemasadia.bukc@bahria.edu.pk (A.S.)

***Abstract*—The Internet is used as a tool to offer people with endless knowledge. It is a global platform which is used for connectivity, communication, and sharing. At almost no cost, an individual can use the Internet to send email messages, update tweets, and Facebook messages to a vast number of people. These messages can also contain unsolicited advertisement which is identified as a spam. The company Twitter too is massively affected by spamming and it is an alarming issue for them. Twitter considers spam as actions that are unsolicited and repeated. These include tweet repetition, and the URLs that lead users to completely unrelated websites. The authors' have worked with twitter's dataset focusing on tweets about "iPhone". It was collected by using an API which was further pre-processed. In this paper, content-based features have been selected that recognize the spamming tweet by using R. Multiple machine learning algorithms were applied to detect spamming tweets: Naive Bayes, Logistic Regression, KNN, Decision Tree, and Support Vector Machine. It was observed that the best performance was achieved by Naive Bayes Algorithm giving an accuracy of 89%.**

***Keywords*—spam detection, twitter, Naive Bayes, machine learning, data analysis, artificial analysis**

I. Introduction

In today's time, there is a deluge of online data which has resulted in the increase of spam data.

There are three perspectives of content based spam detection, such as syntax, semantic and stylistic. The syntax based spam detection predicts the falseness of content by applying different models such as support vector machine or neural network [1].

Our motivation for this research is because spam data is one of the most demanding problem online social networks. A spam can be considered as a message that can be in various formats such as text, links, etc. These spam messages are generally used for malicious purposes such as advertising, hacking, and harassment [2].

Spam messages are viruses hidden in attachments or links in the body of a tweet. These viruses may not be obvious, and often the spammer will try and trick users into opening the link to gain access to user's computer. Spammers tend to post about trending topics. Spam opinion destroy the image of any product or company which is providing a high-quality service while on the other hand it also boosts the image of a company which producing low quality product or services. Identifying the spam tweet or reviews help companies in maintaining online customer attraction towards their services [3].

Tweets in twitter can be rehashed all through the system, a procedure termed as re- tweeting. A Re-tweeted message typically flinches with "RT username", where "@" symbol signifies a reference to one, who initially updates a tweet, and a single tweet contain hashtags (#) to identify certain topics [4]. Spammers in tweet are determined by numerous goals, stated as to spread advertise to produce viruses, phishing, sales, disseminate pornography, or just to negotiate system reputation. This indulges in contamination real time search, not only this but they can also affect statistics accessible by tweet mining tools and consume additional resources from systems and users. Spam wastes human attention. Moreover, Twitter clearly describes that if a tweet is repeated more than once, contains malicious URLs in tweet, too many hashtags and mentions of a spam phrase in a tweet, is considered as spam tweet [5].

Keeping the above Twitter rules in mind; feature selection has been done. The dataset of four thousand tweets were collected for training on different machine learning algorithms. The authors investigated the probability of applying supervised machine learning algorithms (Decision Tree, Naive Bayes, K-Nearest Neighbor, Logistic Regression and Support Vector Machine) to classify spamming tweets. Machine Learning Algorithms were implemented to calculate the accuracy, specificity, sensitivity, and precision. This helped in analyzing which of the algorithm had the highest accuracy level in identifying spam.

II. Literature Review

According to the famous research by Chao Chen of 2017 stated that spam in twitter is denoted as unconstrained tweets comprising malevolent connection which guides casualties to outer destinations containing

Manuscript received August 29, 2022; revised October 13, 2022; accepted October 31, 2022; published March 8, 2023.

doi: 10.12720/jait.14.2.178-184

malware downloads, scams, as well as harms the entire Internet, it was also stated in the month of September 2014, that Internet within New Zealand was softening down because of vast blowout of malware spam downloads. This sort of spam misled clients to connect joins which asserted to contain photographs of Hollywood stars, however its guide clients to download malware to make DDoS attacks possible [6]. As the online world evolves the chances of spam has increased with the increase of data. In 2020, Kouvela [7] done a research on bot detection which is basically spam that can be used for various purposes such as to promote any products. This research resulted in an API development which is known as "Application Programming Interface". A web application is built by using mentioned API in order to identify spam or bots.

In November 2017, Ahmed and Traore *et al.* [8] said that false reviews and false news are related phenomenon as it spreads false statistics. The problem over spam was articulated long time ago, but it quickly become a mounting study area because of plenty generated content of users. Now it is quite easy for anyone to write fake reviews or news on web. In January 2018, according to Walt and Eloff [9], features can likewise be built by consolidating traits from a SMP account, past built highlights, or potentially area learning. A case of a built component is the mix of quantity companions and supporters towards their association as a proportion for contribution to all machine learning algorithms. Features being used by these models are generally expressed as "engineered features", these features are a blend of characteristics called "attributes" and features. Kabakus and Kara [10] look into demonstrated Methods of Detection on Twitter. It expressed about Twitter being most prevalent online networking stages which give an informal community of clients present messages up on 140 characters called as "tweet". Twitter gives clients a chance to segment their messages to be identified including news, big names, occasions, legislative issues. Rendering to Twitter, Twitter's active users are about 313 million months to month and these users post 500 tweets approximately [11]. Twitter gives real-time search figure of what is occurring on the planet with least postponement. Assessment examining administrations can decide about subjects in Twitter which transforms Twitter into an ongoing survey framework. The accomplishment of those administrations totally depends on sifting spammers from authentic clients. Walt and Kara [10] emphasis that different choices were assessed to acquire a dataset of possibly misleading people, given past research, for SMPs. A few analysts utilized information from accessible datasets, similar to pedophiles and radicalism gatherings, to label accordingly.

We have seen the extensive use of Short Message Service for the advertisement of different products which sometimes overloaded our inbox folder. In 2018, Gupta and Bakliwal *et al.* [12] took two different data et of SMS and applied different machine leaning classifiers and both data set's results support CNN (Convolutional Neural Network) for giving high accuracy. In their paper they discussed machine learning techniques for spam detection on SMS. After evaluation of different algorithm, they found that Convolutional Neural Network Classifier achieves the highest accuracy of 99.19% and 98.25%.

In 2020, Govil and Agarwal *et al.* [13] presented a machine learning-based spam detection mechanism. For this purpose, they have used a dataset of approximately 6000 (valid and invalid) emails. They removed all helping verbs from their data set and after that their algorithm will check for the validity of email address that weather it is spam address or not (there is a collection of email addresses from which they never wanted to receive emails). They have incorporated a data structure which will run through Naive Bayes algorithm and as a result the prediction model will be achieved. Most importantly this proposed algorithm detect spams based on patterns. Instead of individual words it will detect the section of spam words in an email. They stated that the application their mechanism will help in terms of getting less number of spam emails.

In 2022, Rodrigues and Fernandes *et al.* [14] contributed their work for spam detection. The primary focus of their work was to detect spam from tweets. After extracting features from tweets, different classifiers were applied such as decision tree, random forest, multinomial na¨ıve Bayes, logistic regression, stochastic gradient descent, support vector machine, logistic regression, random forest, Naive Bayes, and deep learning methods, namely, simple recurrent neural network (RNN) model, long short-term memory (LSTM) model, bidirectional long short-term memory (BiLSTM) model. They have also performed sentiment analysis n tweets by using 1D convolutional neural network (CNN) model. After analyzing all mention models, they came to know that deep learning model (LSTM) has achieved hight acuuracy (98.74%) from all models.

A. Classification Methods

Sun and Lin *et al.* [11] did experiments with several machine learning algorithms such as Naive Bayes, KNN, BLR, GBM, NB and Neural network (NN). This study used tweets data. The input data set also contains URLs because of increase in malicious URLs in tweets. Particularly deep learning algorithms gave about 80% accuracy.

The evaluation of the general procedure depends on an arrangement like measures regularly utilized in Machine Learning and Information Retrieval in the most research paper of 2018 by International Journal of Engineering & Technology. In research of 2018, the evaluation process was in light of an arrangement like set of measures frequently applied in algorithms of Machine Learning and Retrieval of information for classifiers. The precision, accuracy, recall, and F-measure are accounted separately. Every classifier was trained multiple times like almost 10 times, and utilizing this time 9 out of 10 are partitioned as training data. The assessment metrics were evaluated on the basis of average of generated confusion matrixes. It concluded that Naive Bayesian classifier amongst all had the best overall performance in the latest paper of 2018 [15].

B. Twitter Spam Detection by Applying and Utilizing Various Machine Learning Algorithms

The research by S. Nithyanantham, M. Sangeetha, and M. Jayanthi says that twitter spams ordinarily imply tweets containing advancements, drugs arrangements or tweets occupying customers to external dangerous associations containing malware downloads or phishing as well Spams on Twitter can easily affect online social practice, and additionally weakens the security of the web. It further confirmed the utilization of these classifiers proceeding Twitter spam acknowledgment, to validate the importance as a general rule of context. The spam in tweets distinguished within Data Sets are accumulated and explored by using diverse Machine Learning estimations for its execution, soundness and flexibility. The enlightening gathering consists of two features: Account features are, no_lists, no_tweets, account_age, no_follower, no_userfavourites, no_following and content features are, no_char, no_hashing, no_digits, no_retweets, no_usermention, no_url. Many algorithms work over Machine Learning; the algorithms utilized here are Random Forest, K-NN, Stochastic GBM, c5.0 and Naive Bayes. Algorithms chosen are ordinarily utilized in current and scholarly fields. KNN machine learning algorithm is picked, as a result of its accuracy of data tests with genuinely unobtrusive measure of estimations. It grants weighting system for the nearest neighbor's in perspective of their comparability to an unclassified precedent. The probability of a late watched test having a place in class is affected or weighted through the similarity to rest of the models within training set [15].

C. Methods of Machine Learning

By utilizing these machine learning methods, a machine can absorb individually. Thus, no individual intercession is required. These machine learning algorithms utilizes training set as training sets are marked models acquired after physical data evaluation. A paper named "A Spam Transformer Model for SMS Spam Detection" was published in 2021. The primary focus of this paper was to identify spam in SMS (Short Message Service). This research proposed a modified transformer model in order to identify spam in SMS. Results of this research reflects the accuracy of recall, and F1-Score with the values of 98.92%, 0.9451, and 0.9613, respectively [16].

In 2021 Loukas Ilias, Ioanna Roussaki published a paper in which they shared two methodologies based on natural language processing in order to differentiate the actual user from bot. First method used machine learning algorithms and in second method deep learning was used. Based on the input dataset (Social Honeypot Dataset) results of this study revealed that 60% of the users are legitimate while 40% users are bots/spams [17].

III. Design and Methodology

For our Project, Agile methodology is used as there would be numerous changes during the whole period of project development life cycle. Our main objective is to detect spam in tweets and check the accuracy of algorithm with respect to twitter dataset. For the implementation procedure we used RStudio as a software development interface by using R language. R has gotten revolutionary reforms in Big Data Analytics and different parts of data sciences and data analytics as well. R Programming is the best instrument for measurements, analysis of data and machine learning. Fig. 1 shows the conceptual model of this study.

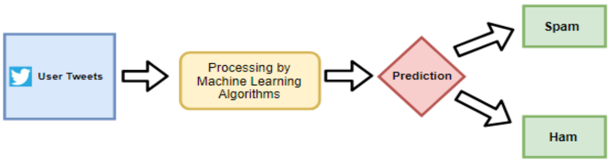


Figure 1. Conceptual model of the spam detection application.

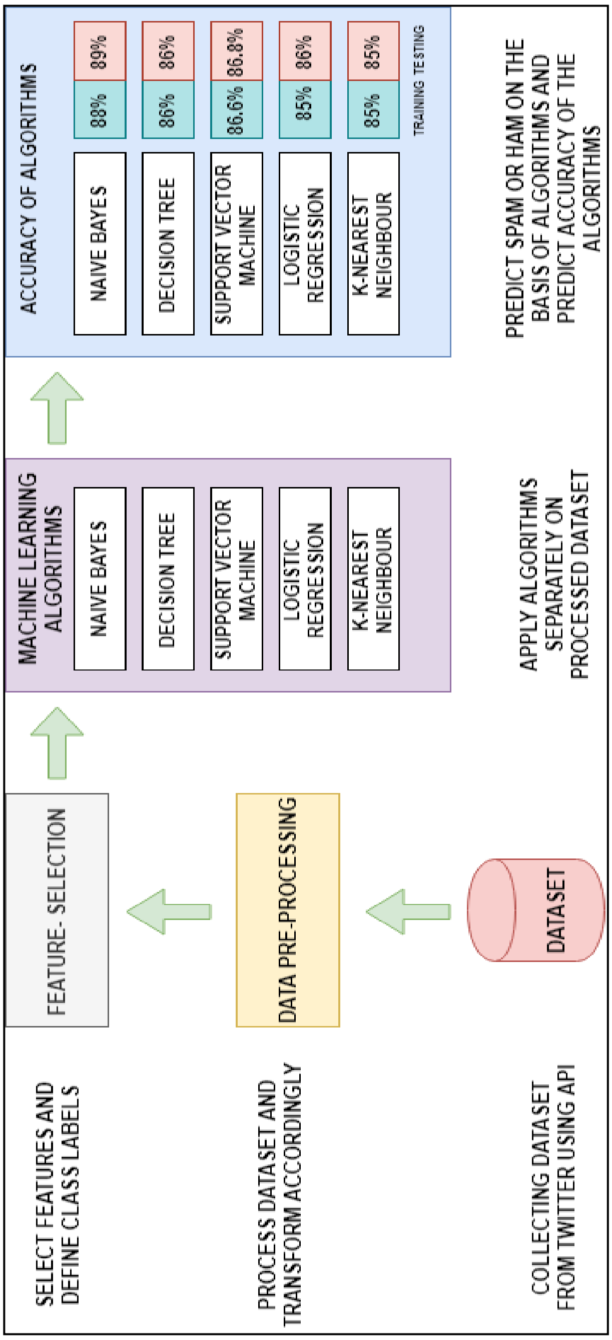


Figure 2. Working of machine learning algorithms.

A. *Implementation*

As shown in Fig. 2, the authors acquired the dataset from twitter using API (TweetR), we created an application in twitter. Intended for the collection of tweets we acquire four keys which are customer key, consumer secret, access token, access token secret. These keys permit us to collect dataset of tweets from twitter according to the twitter rules and policies. After populating twitter's dataset for iPhone, we saved the tweets data into a csv file. In order to process data further, data needs to be clean, for that purpose we performed pre-processing operation on data in which we removed the irrelevant stuffs from text such as punctuations, stopped words and links. Afterwards we selected features from our dataset to make sure that our algorithms give accurate results. So, as far as the feature selection is concerned, we have visited the twitter help Centre website where twitter have defined the rules and regulations that should be followed by users. Following are the features that have been selected from a tweet:

- Hashtags (i.e., #)
- Links (i.e., https://)
- Mentions (i.e., @)
- Spam words
- Repeated tweets

The features that we have chosen from a tweet is: hashtags (i.e., #), links (i.e., https ://), mentions (i.e., @), spam words and repeated tweets to clean our data. By considering the above features that we extracted, the next step is to detect spam so that we can identify that either a tweet is spam or a ham. For this a script is written that checked that how many tweets are repeating and how many tweets contain a spam phrase. Hence, a new dataset is obtained by this process that only consist of the attributes i.e., spam word, repeat, mentions, hashtags, links, and check. Check is the class label that signifies that either a tweet is spam or a ham. This new feature selected dataset is converted into a binary form to be used by the machine learning algorithms.

B. *Experiment*

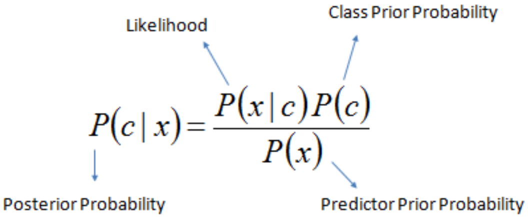
The following machine learning algorithms are used for spam detection:

1) K Nearest Neighbor (KNN)

The KNN stays to be a straightforward algorithm which can store every single accessible case, and which classifies all new cases depending on similar measures, distance functions can be an example. We have used this algorithm for statistical estimation and pattern recognition. While testing the tuple we extracted If *k*=1 (it's a tie 1 vote for spam, 1 vote for ham) and If *k*=3 (2 votes for spam, 1 vote for ham). Thus, the result of Sample test showed that the tuple belongs to SPAM category.

2) Naive Bayes

The Naive Bayes is an exemplary probabilistic machine learning algorithm, expanding on the supposition, that all marked features of a data remains probabilistically independent which are later chosen for possibility analyses. Through Naive Bayes probability formula as shown in Fig. 3, we have calculated probability for every attribute, i.e.,: mentions, repeat, hashtags, links, and spam words.



$$P(c\,|\,x) = \frac{P(x\,|\,c)P(c)}{P(x)}$$

$$P(c\,|\,\mathrm{X}) = P(x_1\,|\,c)\times P(x_2\,|\,c)\times\cdots\times P(x_n\,|\,c)\times P(c)$$

Figure 3. Naive bayes classification formula.

3) Decision tree

As shown in Fig. 4, decision tree solve classification problems simply. This algorithm consists of collection of well-defined questions related to test record attributes. Once we get an answer the follow up question being asked till the final decision made on the record [18].

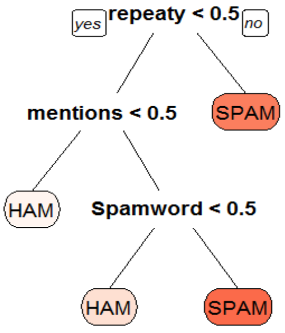


Figure 4. Decision tree.

Decision tree algorithm was used to solve both regression and classification problems. We classified our data set accordingly by the following rules:

- IF repeat=1 THEN tweet is SPAM.
- IF repeat=0 and mentions=0 THEN tweet is HAM.
- IF repeat=0 and mentions=1 and Spam word=0 THEN tweet is HAM.
- IF repeat=0 and mentions=1 and Spam word=1 THEN tweet is SPAM.

4) Logistic regression

Logistic regression takes decision based on threshold value. If the predicted value is greater than the threshold value then it is taken as TRUE or 1 or YES, and on the off chance that if the predicted value is not as much as the threshold value, then it is considered as FALSE or 0 or NO. Its graph typically looks like 'S' because of the curve due to sigmoid function. After taking input dataset of processed tweets, logistic regressions split the data into training and testing datasets. Afterwards a glm model is created which is also known as "Generalized linear model". The summary of model is formed after the creation of glm model which shows the significance of data, i.e., the accuracy of data.

5) *Support Vector Machine*

The Support Vector Machine (SVM) algorithm figures out how to recognize the two categories dependent on a training set of records that contains labelled models from the two categories. It is mostly utilized in grouping issues such as classification problems. For SVM, we plot every piece of information as a point inside the n-dimensional space (where n is said to be the number of highlighted features) together with the estimation of each element in presence with the estimation of a definite coordinate. At this point, we have formulated classification by concluding the separation of two classes.

IV. Results and Discussion

In this research it can be concluded that multiple machine learning algorithms were thoroughly tested to correctly identify ham and spam tweets. Naive Bayes algorithm outperforms other algorithms in the detection of spam content. Fig. 5, shows the comparison of the performance of each machine learning algorithm.

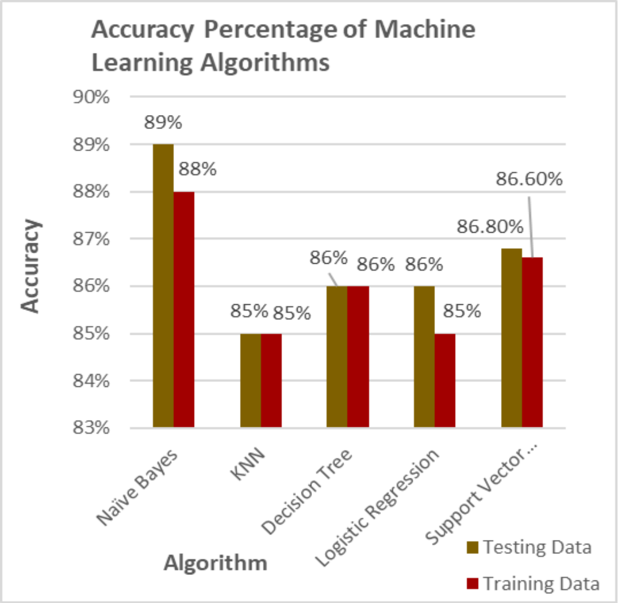
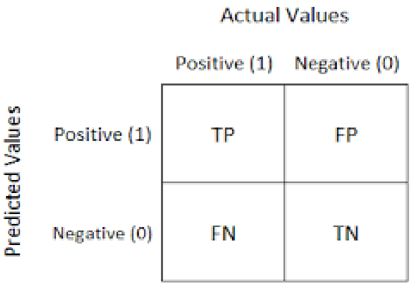


Figure 5. Accuracy results of ML algorithms.

TABLE I: Combined Confusion Matrix of All Implemented Algorithms

P1 For Training data			P2 For Testing data		
Naive Bayes					
P1	HAM	SPAM	P2	HAM	SPAM
HAM	1069	11	HAM	257	3
SPAM	368	1752	SPAM	85	455
KNN					
P1	HAM	SPAM	P2	HAM	SPAM
HAM	1166	79	HAM	486	48
SPAM	317	1237	SPAM	123	544
Decision Tree					
P1	HAM	SPAM	P2	HAM	SPAM
HAM	1244	374	HAM	531	158
SPAM	2	1181	SPAM	2	508
Logistic Regression					
P1	HAM	SPAM	P2	HAM	SPAM
HAM	1086	112	HAM	530	51
SPAM	284	1186	SPAM	133	618
SVM					
P1	HAM	SPAM	P2	HAM	SPAM
HAM	1244	371	HAM	531	156
SPAM	2	1184	SPAM	2	510

Above Confusion matrix in Table I determine true positive, true negative, false positive and false negative describe below in Fig. 6:



Predicted Values	Actual Values: Positive (1)	Actual Values: Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Figure 6. Confusion matrix.

True positive: detects the condition when condition is present.

False positive: detects the condition when the condition is absent.

False negative: does not detect the condition when condition present.

True negative: does not detect the condition when the condition is absent.

For our results we also calculated the accuracy, precision (fraction of all positive identification was actually correct), specificity (it is the measurement of the proportion of actual negatives that are accurately distinguished known as the true negative rate) and sensitivity (as true positive rate, or recall. It quantifies the extent of true positives that are effective.

$$Accuracy = \frac{TN + TP}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Table II shows the comparison amongst accuracies of machine learning algorithms, i.e., Naive Bayes, Decision Tree, Support Vector Machine, Logistic Regression, K Nearest Neighbor, that Naive Bayes with 89% accuracy has the highest accuracy among all. It correctly predicted the output for testing data of twitter for spam or ham tweets

Combined results where Decision Tree and K-Nearest Neighbor has 86% and 85% and Logistic Regression and Support Vector Machine has 86% and 86.8% accuracies respectively.

TABLE II. ACCURACY COMPARISON OF ML ALGORITHM

For Training data		**For Testing data**	
Naive Bayes			
Accuracy	0.88	Accuracy	0.89
Sensitivity	0.743	Sensitivity	0.751
Specificity	0.993	Specificity	0.993
Precision	0.989	Precision	0.988
KNN			
Accuracy	0.85	Accuracy	0.85
Sensitivity	0.786	Sensitivity	0.798
Specificity	0.931	Specificity	0.918
Precision	0.936	Precision	0.910
Decision Tree			
Accuracy	0.86	Accuracy	0.86
Sensitivity	0.998	Sensitivity	0.996
Specificity	0.759	Specificity	0.762
Precision	0.768	Precision	0.770
Logistic Regression			
Accuracy	0.85	Accuracy	0.86
Sensitivity	0.79	Sensitivity	0.79
Specificity	0.91	Specificity	0.92
Precision	0.90	Precision	0.91
SVM			
Accuracy	0.866	Accuracy	0.868
Sensitivity	0.998	Sensitivity	0.996
Specificity	0.761	Specificity	0.765
Precision	0.770	Precision	0.772

V. CONCLUSION AND FUTURE WORK

Regardless of the nature of spam content, it can still be detected and deleted automatically. This is only possible by an exemplary machine learning algorithm and for this research Naive Bayes outperformed every other machine learning algorithm. Hence the authors based on the highest accuracy achieved i.e., 89% on their dataset, culminate this research with Naive Bayes being the best spam detecting algorithm. In future, the authors intend to further evaluate this dataset using deep learning algorithms i.e., Convolutional Neural Networks.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

AUTHOR CONTRIBUTIONS

Azeema and Ammarah collected the data through Twitter API and pre-processed the data. They tested the working of the support vector machine on the dataset. Fatima and Reema tested the performance of Naive Bayes algorithm, KNN, and Decision Tree on the filtered dataset. Amna tested the Logistic Regression Algorithm. Azeema etal., performed the comparative analysis based on their research findings.

REFERENCES

[1] A. Mewada and R. K. Dewang, “A comprehensive survey of various methods in opinion spam detection,” *Multimedia Tools and Applications*, pp. 1–41, 2022.

[2] Z. F., Sokhangoee and A. Rezapour, “A novel approach for spam detection based on association rule mining and genetic algorithm,” *Computers & Electrical Engineering*, vol. 97, p. 107655, 2022.

[3] R. M. Saeed, S. Rady, and T. F. Gharib, “An ensemble approach for spam detection in Arabic opinion texts,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 1, pp. 1407–1416, 2022.

[4] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, “Detecting spammers on twitter,” in *Proc. Collaboration, Electronic Messaging, Anti-abuse and Spam Conference (CEAS)*, vol. 6, no. 2010, 2010, p. 12.

[5] Twitter help center. [Online]. Available: https://help.twitter.com/en

[6] C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou, and G. Min, “Statistical features-based real-time detection of drifted twitter spam,” in *Proc. IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, 2016, pp. 914–925.

[7] M. Kouvela, “Bot detective: Explainable bot detection in twitter,” A thesis submitted in fulfillment of the requirements for the degree of Master of Data & Web Science, 2020.

[8] H. Ahmed, I. Traore, and S. Saad, “Detecting opinion spams and fake news using text classification,” *Security and Privacy*, vol. 1, no. 1, p. e9, 2018.

[9] E. Van Der Walt and J. Eloff, “Using machine learning to detect fake identities: Bots vs humans,” *IEEE Access*, vol. 6, pp. 6540–6549, 2018.

[10] A. T. Kabakus and R. Kara, “A survey of spam detection methods on twitter,” *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 3, 2017.

[11] N. Sun, G. Lin, J. Qiu, and P. Rimba, “Near real-time twitter spam detection with machine learning techniques,” *International Journal of Computers and Applications*, vol. 44, no. 4, pp. 338–348, 2022.

[12] M. Gupta, A. Bakliwal, S. Agarwal, and P. Mehndiratta, “A comparative study of spam SMS detection using machine learning classifiers,” in *Proc. 2018 IEEE Eleventh International Conference on Contemporary Computing (IC3)*, 2018, pp. 1–7.

[13] N. Govil, K. Agarwal, A. Bansal, and A. Varshney, “A machine learning based spam detection mechanism,” in *Proc. 2020 IEEE Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 954–957.

[14] A. P. Rodrigues, R. Fernandes, A. Shetty, K. Lakshmanna, and R. M. Shafi, “Real-time twitter spam detection and sentiment analysis using machine learning and deep learning techniques,” *Computational Intelligence and Neuroscience*, 2022.

[15] M. Sangeetha, S. Nithyanantham, and M. Jayanthi, “Comparison of twitter spam detection using various machine learning algorithms,” *International Journal of Engineering & Technology*, vol. 7, no. 1–3, pp. 61–65, 2017.

[16] X. Liu, H. Lu, and A. Nayak, “A spam transformer model for SMS spam detection,” *IEEE Access*, vol. 9, pp. 80253–80263, 2021.

[17] L. Ilias and I. Roussaki, “Detecting malicious activity in Twitter using deep learning techniques,” *Applied Soft Computing*, vol. 107, p. 107360, 2021.

[18] N. Ahmed, R. Amin, H. Aldabbas, D. Koundal, B. Alouffi, and T. Shah, “Machine learning techniques for spam detection in email and IoT platforms: Analysis and research challenges,” *Security and Communication Networks*, 2022.

Azeema Sadia received her BSCS degree in computer science in 2013 from Bahria University Karachi Campus, Pakistan and MSCS degree in Computer Science in 2016 from the National University of Computer and Emerging Sciences-FAST Karachi, Pakistan. She is working as senior lecturer at Bahria university Karachi campus. She has about 8 years of teaching experience (undergraduate classes of computer science). The author’s main interests include applications of data analysis such as sentiment analysis or opinion mining.

Fatima Bashir is currently working as a senior lecturer in Bahria University with the six year of experience teaching undergraduate level. She has done her bachelor's in computer science & information technology form NED University of Engineering and Technology and completed her master in 2017 with the thesis based on machine learning. The authors' area of research includes modeling and simulation, prediction analysis and sentiment analysis. Moreover, she is trying to contribute her research in mentioned areas and have also publications.

Amna Bashir has completed her masters in 2021 as a software engineer from renowned university of Pakistan and have four-year experience at industry as a software quality assurance engineer. Now she is teaching as a lecturer in Sir Syed University of Engineering & Technology. She is very passionate in teaching programing courses including object oriented programming, database management system and web programming. Her research interest includes opinion mining and System development.

Engr. Ammarah Khalid received her masters and bachelor's degree in software engineering, both from Bahria University Karachi Campus, Pakistan. She is currently serving as senior lecturer at Bahria University Karachi Campus and she is the cluster head of Software Engineering Dept. of BUKC. She has about 8 years of teaching experience (undergraduate classes of Software Engineering) and 2 years of cooperate experience as a software developer. The author's main interests include image visualization, human computer interaction and natural language processing.

Engr. Reema Qaiser Khan has received her advanced master's degree in artificial intelligence from Katholieke Universiteit Leuven Belgium in 2021. Her thesis was on natural language processing. She also has a master's degree in software engineering and a bachelor's degree in computer engineering from renowned Pakistani Universities. She currently has 9 publications with 24 Citations and h-index 3. She is currently working as a senior lecturer at Bahria University Karachi Campus in the Computer Science Department. The author's prime interest lies in AI health care and cognitive AI.