International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Mar-24)

http://dx.doi.org/10.12785/ijcds/150174

Detection of Roman Urdu fraud/spam SMS in Pakistan Using Machine Learning

Muhammad Ayaz[1], Sarwat Nizamani[2], Aftab Ahmed Chandio[3] and Kirshan Kumar Luhana[4]

[1,3] *Institute of Mathematics and Computer Science - University of Sindh, Jamshoro, Pakistan*
[2] *University of Sindh, Mirpurkhas Campus, Pakistan*
[4] *University of Sindh, Laar Campus, Badin Pakistan*

Received:18 May 2023, Revised:12 Feb. 2024, Accepted:24 Feb. 2024, Published:1 Mar. 2024

Abstract: Over the past few years, mobile devices and their services have become widely used around the world. Almost everyone uses the Text Messaging Service (SMS) for communication purposes because it is easy to use and inexpensive. When a person tries to deceive another for the sake of profit (material or money), it is known as Fraud. Through SMS fraud, fraudsters often adopt various strategies to make their messages look credible and legitimate. Various popular organizations use SMS services to advertise their products and send messages to individuals about their services. As a result, one receives many junk messages. Spam message is a message sent to any user who does not want to have it on their phone. Spam or fraudulent messages can be threatening and can sometimes cause financial and confidential data loss. In Pakistan, messages are sent in English and Urdu (Pakistani national language) but most messages are sent using Roman Urdu (Urdu written using Latin / English characters). This research compares the strategies and algorithms used in the literature to detect spam / fraudulent messages written in English or in any local language such as Roman Urdu. The study also suggests a new way to detect fraudulent messages written directly in Roman Urdu. In the fraud detection process, three different monitoring machine learning classifiers are used in this study namely Support Vector Machine (SVM), Naïve Bayes (NB) and Decision Tree (J48). After using the model, we found that SVM performed better than the other two classifiers with 99.42% accuracy.

Keywords: spam, fraud, Roman Urdu, machine learning

1. Introduction

Short Message Service (SMS) is increasing day by day due to its low price and ease of use. SMS is used for basic communication between individuals, but there are so many companies and organizations that use SMS service for advertisement purpose. The use of electronic messaging system to send uninvited mass messages mainly for advertisement motive is known as spam. Creating spam with an SMS containing a marketing theme is known as SMS Spam [1]. When one tries to deceive another person to get some benefit (items of value or money), it is known as Fraud [2]. Through SMS fraud, fraudsters adapt different techniques to make their messages look like trustworthy and legitimate. Fraud messages and spam messages are two different categories of SMS. There has been lots of efforts done to deal with spam and fraud; mainly e-mail spam/fraud detection in English language is done by so many researchers in literature (which will be discussed in 'Literature Review' section of this paper). Fraud SMS are further divided into three different categories [2]

- Lottery
- Damsel in Distress
- Steal Credentials

In a "Lottery" fraud, a fraudster sends a text message to an individual mentioning that he has won a large sum of money by a fortune-telling or other scheme and that person must contact a certain number to obtain his or her winning prize. When individual calls to fraudster, he convinces him to pay some amount to get his or her prize. Individual makes the payment, but never gets his prize. In a "Damsel in Distress" fraud, the impostor proves to be a poor young lady, desperately in need of help. The fraudster requests to send him some cell phone credits and promises to return them back. In a "Steal Credentials" fraud, fraudster shows himself as Bank or some other organization and tries to steal the user credentials by saying that your ATM card has been blocked or some services have been suspended. In order to reactivate them, you need to provide your personal information i.e., password, PIN etc. [2]. Urdu is the national language of Pakistan and people are more comfortable communicating in their national language. That is why Roman Urdu (Urdu written

E-mail address: ayaz.khatri@usindh.edu.pk, sarwat@usindh.edu.pk, chandio.aftab@usindh.edu.pk,kirshan.luhano@usindh.edu.pk

in Latin/English alphabets) is widely used as a means of SMS communication among Pakistanis. Most of the research studies in the literature, to detect the spam/fraud, are based on English language and there are only a few researches that are based on local languages such as Roman Urdu. This research study elaborates the methodologies of previous studies and suggests a new strategy to spam/fraud detection problem. The structure of this research study is as follows: it starts with the introduction and discussion of the literature review, followed by a summary of the problem statement. After that, proposed methodology is discussed in details and the results of various classifiers are then displayed. Finally, the study concludes with a summary of its findings and recommendations for further research.

2. Literature review

This section elaborates the previous studies in literature that detect the spam/fraud in English language or in Roman Urdu. Table I gives an overview of these studies. Mujtaba, G., & Yasin, M. [1] collected a total of 6600 messages from different users who volunteered to contribute in dataset. They extracted four features from each message. In first feature, all characters of each message were count observing that ham (regular) messages have less characters than spam messages. If a message has less characters than it is more likely to be ham message. In second feature, spam words were matched with a list of most occurring words in spam messages. If a match was found than this feature was set to 1 and more likely to be a spam message. In third feature, the combination of words is checked against most occurring combinations in spam messages i.e., 'Activate Now', 'Buy Now'. If some of these combinations were found in message than it is more likely to be a spam message. In fourth feature, SMS class/label was defined, 1 indicates spam message; 0 indicates ham message. Waikato Environment for Knowledge Analysis (WEKA) [3] was used for classification in this study. Various classifiers of supervised learning in WEKA were utilized which are Naive Bayes [4], Neural Network (Multilayer Perceptron) [5] , C4.5 Decision Tree (J48) [6] and their results were compared. Each algorithm was given 2244 instances and 66% of data were used for training purpose and the rest was used for testing purpose. Naïve Bayes gave the accuracy of 92.953%. Multi-Layer Perceptron gave the accuracy of 89.3048% and C4.5 classifier gave the accuracy of 89.3048%. Naïve Bayes algorithm produced more accuracy than other two algorithm for the given dataset. Karami, A., & Zhou, L. [7] extracted two different kinds of features from messages, first kind of features are SMS-Specific (SMSS) features and second ones are Linguistic Inquiry and Word Count (LIWC) features. In SMSS features, the rates of different keywords are extracted such as rate of URL, rate of Spam Word (SW), rate of Unique Words (UW) and others. In LIWC features, the categories of text semantics are considered such as score of punctuations, score of pronouns, score of verbs and others. They used a publicly available dataset of 5574 messages. 86.6% of which were non-spam messages and 13.4% messages were spam messages. To classify messages as spam and non-spam, WEKA was used. 40 different algorithms of Supervised Machine Learning were utilized and their results were compared in this research. Most of the algorithms showed an accuracy of 92% to 98%. Support Vector Machine (SVM) [8] and Random Forest [9] gave the best performance among all. Almeida, T. A. et al. [10] collected messages from various websites and combined into a collective larger dataset for this study. Dataset has a total of 5574 messages, 86.60% were Ham messages and rest were spam messages. Because data were collected from different sources, there was a probability of duplication of data. Duplicate messages were eliminated. To apply Machine Learning algorithms, WEKA was used. The algorithms that were used in this study were Naïve Bayes (NB), SVM, Minimum Description Length (MDL) [11], K-Nearest Neighbor (KNN) [12] , C4.5 and PART [13] . SVM gave the highest accuracy of 97.65% and performed well among other algorithms. PART gave the accuracy of 97.50%, MDL gave the accuracy of 96.26%, C4.5 gave the accuracy of 95.00%, 1NN gave the accuracy of 92.70%, Naïve Bayes gave the accuracy of 92.05% and 3NN gave the accuracy of 90.10%. Khan, M. S. et al. [14] used an android app to collect the messages from individuals. Users of this app were asked to willingly contribute inbox messages and label them according to their choice (spam/non-spam). The dataset contained 8107 messages after preprocessing which includes spam and non-spam messages both. Firstly, dataset was transformed into lower-case and stop words were removed to increase efficiency of computational resources. Stop words are those words which cannot identify if a message is spam [15]. 90% dataset was used for training purpose and 10% dataset was used for testing purpose. Support Vector Machine Model from Python's Scikit library was used for this study. The model was trained to classify the messages based on how individual users have labeled the messages (spam/non-spam). The results of this study gave an accuracy of 96.8Mehmood, K. et al. [16] used a manual dataset of 8449 messages. In Data Pre-Processing phase, all non-alphanumeric characters were removed because these characters have nothing to do with spam filtering. Very short messages (less than six characters) were also removed because these messages do not provide much information. Duplicate messages or messages having same context were also removed. Finally, all messages were transformed to lowercase letters. In Data preparation for classification phase, all messages were labelled spam/ham by domain experts. After that dataset was converted to the .arff format that is acceptable by WEKA. The algorithms that were used in this research do not work with text instances that is why all text instances were converted to vector format. To do this "StringToWordVector" filter of WEKA was used. In Spam Filtering (Data Classification) phase, WEKA Tool is used for classification. Different algorithms such as Naive Bayes Multinomial, DMNBText [17], LibSVM, LibLinear [18] and Sequential Minimal Optimization (SMO) were used for the purpose of spam filtering and their results were compared. Results were produced using 10-fold cross validation for all algorithms. Accuracy compares

the correctly classified instances with incorrectly classified instances and gives results in percentage. SMO gave the highest accuracy of 93.3%. DMNBText gave the accuracy of 92.74%. Naïve Bayes Multinomial gave the accuracy of 92.22%. LibLinear gave the accuracy of 91.42%. SVM gave the accuracy of 88.42%. All algorithms except SVM gave the accuracy more than 90%. Afzal, H., & Mehmood, K. [19] used WEKA for classification of tweets. As many as 1463 different tweets were collected through Twitter Streaming API. After collection of data, tweets are passed to a JAVA program for pre-processing such as removal of non-alphanumeric characters and removal of tweets that were less than 6 characters long. After that tweets are labelled as spam or ham by experts and dataset was converted to .arff format because this file format is supported by WEKA. For classification of tweets "10 Folds Cross Validation" rule is used. The algorithms that are used in this study does not work with strings so the text was converted to vector format with WEKA filter "StringToWordVector". Five different algorithms, which are Naive Bayes Multinomial, DMNBText, LibSVM, LibLinear and J48, were used for classification of tweets and their results were compared. Naïve Bayes Multinomial gave the highest accuracy of 95.42%. The accuracy of DMNBText was 95.12%. LibLinear gave the accuracy of 94.60%. The accuracy of J48 was 91.38% but it took 11.33 seconds which is not a feasible time for smartphone's environment. LibSVM accuracy was 70.88% which was worst of all algorithms. Nizamani et al. [20] used WEKA tool to detect suspicious emails that have any content regarding terrorist attack in near future. They used a manual dataset for classification of emails. 45% of dataset contained the suspicious emails and rest of the emails were non-suspicious. Four classifiers with feature selection abilities were utilized. Logistic regression [21], decision tree (ID3) [22], Naïve Bayes and SVM were used for classification. Logistic regression gave an accuracy of 83.92%. Decision tree (ID3) also gave an accuracy of 83.92%. Accuracy of SVM was 80.35% and Naïve Bayes underperformed with an accuracy of 78.57In another study, Nizamani et al. [23] detected the fraudulent emails by using special feature selection. For this task they utilized the WEKA tool and used different classification algorithms such as Naïve Bayes, SVM, J48 and Cluster Based Classification Model (CCM) [24]. They used a dataset having 8000 emails, 50% emails were fraudulent emails and others were normal emails. Initial features were extracted using TF-IDF [25] scheme and other features were added manually. The results were calculated using 10-fold cross validation. In this study, the highest accuracy achieved was 96%.

3. Problem statement

Urdu is the national language of Pakistan and people feel more comfort when they are communicating in their national language. That's why Roman Urdu is mostly used as medium of communication via SMS in Pakistan [2]. Due to rapid use of SMS, many fraudsters use SMS services to get illegal benefits from innocent people (categories of fraud messages are already described above). There is a dire need of such a system/model that detects the fraud messages, specifically typed in Roman Urdu, as soon as they are received on one's phone. Such a system/model can save many individuals from financial or confidential data loss. Many researchers [16] [1] [14] [19] [7] [10] [20] [23] have done efforts to deal with spam/fraud in literature. English language texts have been the main focus of recent research studies in the field of SMS fraud detection. These studies make use of cutting-edge natural language processing (NLP) methods and machine learning algorithms. Roman Urdu being used as a target language for fraud SMS detection, makes a substantial contribution to the literature. This change not only broadens the scope of the research but also takes on a critical issue in areas where Roman Urdu or any other local language is frequently used for text messaging. In Pakistan, the percentage of frauds are increasing day by day. Fraudsters trap the individuals by using different and new tactics and get illegal benefits from them. As a result, the individual suffers from financial and sometimes mental health loss [2]. To the best of our knowledge, there is not any effort done to detect fraud messages that are typed in Roman Urdu and there is not any publicly available dataset that contains Roman Urdu fraud message.

4. Proposed methodology

This section describes the proposed methodology of this study to detect the fraud messages that are sent in Romanized Urdu. For this study WEKA tool is used because WEKA's user-friendly interface, visualization tools, large algorithm library, integrated data preprocessing and evaluation capabilities and open-source nature make it an ideal tool for implementing machine learning classifiers like SVM, Naïve Bayes, and Decision Trees. Fig. 1 gives the overview of different phases of the proposed methodology.

A. Data collection

In data collection phase, the Romanized fraud messages are collected from different sources. To the best of our knowledge, there is not any publicly available dataset of Romanized Urdu fraud messages that's why a new manual dataset is created for this study. To collect the data, a WhatsApp group is created and some volunteers are added into it who willingly contribute their messages in that group. All volunteers shared the Romanized fraud messages as well as normal messages. One of the problems that we encountered while collecting data is that many fraud messages were similar or with only a minor difference among them. To overcome this issue, we asked volunteers to manually type some fraud messages in Roman Urdu and contribute in our dataset. By asking multiple volunteers to type messages, we also ensured that messages are coming from different minds of people because fraud message can be written by anyone with multiple variants in writing. After performing some operations on the dataset instances (these operations are discussed in later sections of this study), our final dataset contains 1050 messages. Some of the fraud messages from our dataset with English translation is shown

TABLE I. LITERATURE REVIEW

#	Reference	Language	Tools	Dataset	Algorithms with accuracy	
01	Mujtaba, G. et al. [1]	English	WEKA	6600 Messages	Naive Bayes: Multilayer Perceptron: J48:	92.953% 89.3048% 89.3048%
02	Karami, A. et al. [7]	English	WEKA	5574 Messages	Different Algorithms, such as Random Forest, Naïve Bayes, SVM and 40 other algorithms. Most algorithms gave accuracy 92% to 98%	
03	Almeida, T. A. et al. [10]	English	WEKA	5574 Messages	SVM: Naïve Bayes: MDL: 1NN: 3NN: C4.5: PART:	97.64% 92.05% 96.26% 92.72% 90.10% 95.00% 97.50%
04	Khan, M. S. et al. [4]	Roman Urdu	Scikit Library of Python	8107 Messages	SVM:	96.8%
05	Mehmood, K. et al. [16]	Roman Urdu	WEKA	8449 Messages	Naive Bayes Multinomial: DMNBText: LibSVM: LibLinear: SMO:	92.22% 92.74% 88.42%. 91.42% 93.3%
06	Afzal, H. et al. [19]	Roman Urdu	WEKA	1463 Tweets	Naive Bayes Multinomial: DMNBText: LibSVM: LibLinear: J48:	95.42% 95.12% 70.88% 94.60% 91.38%
07	Nizamani et al. [20]	English	WEKA	Not Mentioned	Logistic Regression: ID3: SVM: Naïve Bayes:	83.92% 83.92% 80.35% 78.57%
08	Nizamani et al. [23]	English	WEKA	8000 E-mails	Naïve Bayes: SVM: J48: CCM:	Highest accuracy achieved was 96%

in Table II. To maintain confidentiality, complete mobile numbers are not shown.

B. Data cleaning

In data cleaning phase, all duplicate messages are removed and only distinct messages are part of the dataset. All special characters, links, emoji's, punctuation symbols or anything, that has nothing to do with fraud detection, are removed from each instance of dataset. Microsoft Excel's Power Query is used to remove special characters, links, emoji's, and punctuation symbols. The following expression is used to achieve the desired results: Text.Select([Data with Special Characters],"A"..."z","0"..."9"," ") This expression is used in Excel's Power Query Editor to clean and filter text data by selecting only spaces and alphanumeric characters while removing special characters, links, emoji's, and punctuation symbols. All messages that have less than four words are also removed because messages having less than four words are not long enough to detect fraud. After completing data cleaning phase, dataset contained 978 messages.

C. Data Pre-Processing

In data preprocessing phase, all the messages are labeled manually as fraud/normal messages. Initially each message is a whole string, and to perform classification, string message needs to be converted in vector format. WEKA does not work with string instances that is why all messages were converted to vector format (that is acceptable by WEKA) by using StringToWordVector function of WEKA.

D. Feature extraction and weighting

Feature extraction is one of the major steps for classification task. For text classification, usually words are considered as features. Initially, the features which do not contribute in identifying the fraud SMS are removed, later the weights are calculated for remaining words. These two steps are further described below.

1) Stop words removal

Stop words are the words that has nothing to do in classification process. By removing stop words, we can save space and enhance the efficiency of our model [15]. The examples of stop words in English language are: "if", "and", "or", "the" etc. Because all instances in our dataset are in Roman Urdu and there is not any dataset available of Roman Urdu stop words so we created a manual stop word text file which contains the Roman Urdu stop words. Some of the examples of Roman Urdu stop words are: "aur", "par", "lekin" etc.

TF-IDF weighting

TF-IDF [25] is an acronym of "Term Frequency-Inverse Document Frequency". In this method, weights are assigned to each word according to its importance in identifying a specific type of message. For instance, if any word which appears only in fraud SMS, that word will be assigned high weightage. On the other hand, if a word appears many times in both types of SMS i.e., fraud and regular, that will be given low weightage.

E. Model training

In model training phase, three classifiers of supervised machine learning are used to classify the normal/fraud messages which are SVM, Naïve Bayes and Decision Tree (J48). These three classifiers are chosen for this study because they are widely used in text classification problems in the literature and have given outclass results and performance than other classifiers. Support Vector Machine (SVM) looks for a hyperplane that maximizes the margin between the two classes while effectively discriminating between normal/fraudulent messages and legitimate data. It works well in high-dimensional spaces and uses kernel functions to handle non-linearity. Based on the Bayes theorem and the assumption of feature independence, Naïve Bayes (NB) uses probabilistic classification. It determines the likelihood that a message falls within the normal/fraud category and bases its judgement on that likelihood. Decision Tree (J48) creates a tree-like structure with each

TABLE II. Roman Urdu fraud messages with English translation

Roman Urdu	English Translation
Dear Customer! Apke bank account ki services expire ho chuki hain. Services ko dobara free me active karne k liye abhi call back karain or 1000 ka free bonus hasil karain. Ye offer sirf 3 din tak k liye hai.	Dear Customer! Your bank account services have been expired. To re-activate services for free please call back and get a bonus of 1000 rupees. This offer is valid for three days only.
JEETO PAKISTAN lucky draw me aap ne hissa liya. Ap ko mubarakbad di jati hai k ap 50000 ki raqam jeet chuke hain. Abhi is number par contact karain 03xxxxxxxxx.	You participated in JEETO PAKISTAN lucky draw. Many Congratulations to you! You have won 50000 cash prize. Contact on this number now 03xxxxxxxxx.
apko khuda ka wasta hai plz mere is number per 100 ka load karwa do me is waqt hospital me hu or mere shohar ki tabiat bohot khrab hai agar yaqin nahi to call me main bat karungi.	For God's sake kindly send me 100 rupees load on my number. I am in hospital right now and my husband is in very serious condition. If you don't believe me then call me, I will explain.

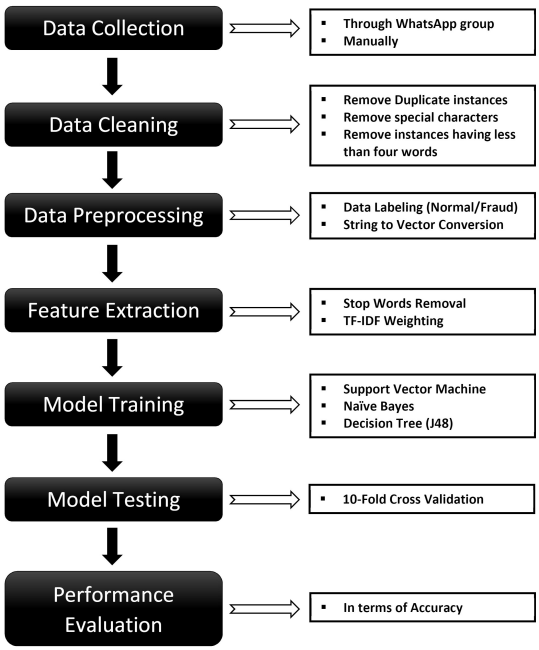

Figure 1. Proposed Methodology

branch representing a decision rule based on feature values. It divides the data into subsets iteratively before classifying messages as normal/fraud. Additional information is given below regarding the functioning of these classifiers.

1) Support Vector Machine (SVM)

SVM is advanced and most commonly used algorithm in the field of text classification[26]. It provides more accurate results when working with large and high dimensional datasets. Moreover, this algorithm can alter the non-linearly separable data into linearly separable data [8] . With the help of non-linear mapping, this algorithm converts the original training data into a higher dimension. It creates a decision boundary, normally called hyperplane, to separate one class from another. The coordinates of this hyperplane are found using support vectors (essential attributes of training data)

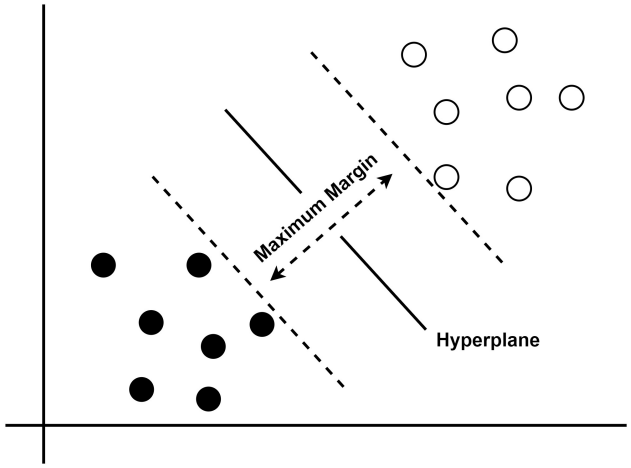

Figure 2. Support vector machine

and the margin defined by those support vectors.

In Fig. 2, simple functionality of SVM is shown, where black circles data belongs to one class and white circle data belongs to another class. To separate the data of one class from other class, a hyperplane (straight line) is created. There can be multiple hyperplanes for this data, the algorithms must choose the best hyperplane that is also called the maximum marginal hyperplane (MMH).

2) Naïve Bayes (NB)

NB is a powerful algorithm which is based on Bayesian theorem. This algorithm is also used for classification of text. It works by calculating the probabilities of features for each group and anticipates a particular class for any given instance [4]. The Bayes Theorem finds opportunities of an event to happen when considering the possibility that another event has already taken place. The Bayes theorem is mathematically defined as the following equation:

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)} \quad (1)$$

In equation 1, A and B are events and P(B) ≠ 0

- Basically, we are trying to find opportunities for event A, as event B is true. Event B is also referred to as evidence.
- P(A—B) is a posterior probability of the event A when event B is already true, i.e., the probability of an event after the identification of evidence.
- P(B—A) is likelihood probability of event B when event A is true.
- P(A) is the prior probability of the event A (pre-probability, i.e., pre-event probability).
- P(B) is the Marginal probability of event B.

3) Decision tree (J48)

J48 is another well-known classification algorithm which is commonly used for its simplicity and inductive nature. In WEKA, J48 is the implementation of C4.5 which is another popular decision tree algorithm [6]. The decision tree uses tree representation to solve the problem where each leaf node corresponds to the class label and the attributes are represented as the internal node of the tree. The major problem in decision tree is to identify the attribute for the root node in each level[27]. For this purpose, attribute selection process 'Information Gain' is normally used. When we use a node in the decision tree to split training instances into smaller sets the entropy changes. The measurement of uncertainty of random a variable is known as entropy. Here is the mathematical formula for entropy in decision tree:

$$Entrophy(S) = \sum_{i=1}^{c} -pi \log 2pi \tag{2}$$

In equation 2, pi is the frequent probability of a class i. Information Gain is the measure of this change in entropy. Here is the mathematical formula of Information Gain in decision tree:

$$Gain(S, A) = Entrophy(S) - \sum_{ve\ Values(A)} \frac{|S\,v|}{S} Entrophy(S\,v) \tag{3}$$

In equation 4, S is a set of instances, A is an attribute, Sv is the subset of S with A = v, and Values (A) is the set of all possible values of A.

4) Model testing

In model testing phase, 10-fold cross validation testing technique is used. In this process, the dataset is split-up into ten subsets and the algorithm completes in ten rounds. In each round, nine subsets are used for training purposes and one distinct subset is used for testing. After completing all ten rounds the average accuracy, gained from all ten rounds, is returned.

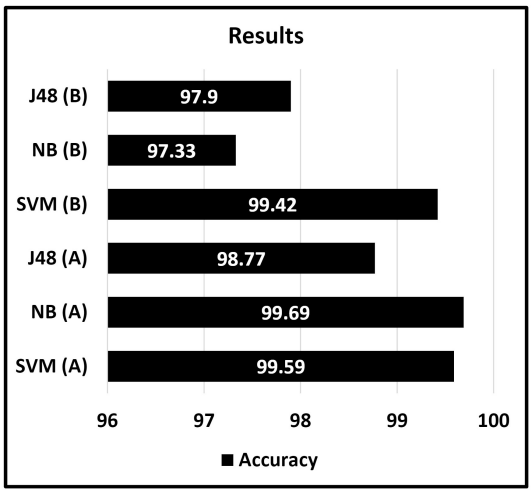

Figure 3. Results

5) Performance evaluation

Results of this model is given in terms of accuracy which is the ratio of the correctly classified messages as compared to the incorrectly classified messages. Here is the formula of accuracy:

$$Accuracy = \frac{Correctly\ classified\ messages}{Overall\ messages} \tag{4}$$

Complete result details for each algorithm are given in the later section of this paper.

5. Results and discussion

This section explains the results of each classifier that is used in this study in terms of accuracy. Initially:

- SVM gave an accuracy of 99.59
- NB gave an accuracy of 99.69
- J48 gave an accuracy of 98.77

Above results are initial results when there are not very much prominent terms of fraud messages in normal messages. These are the results that are generated from the dataset that was collected from different users of mobile phone who voluntarily contributed in the dataset. The dataset also contained the normal chatting messages of users which do not have prominent terms of fraud messages. In order to train the model efficiently, we manually added 72 more normal messages having prominent terms of fraud messages i.e., "mubarak ho", "call back", "madad", "jeeto Pakistan", etc. Now our final dataset contains 1050 messages having 484 fraud messages and 566 normal messages.

After executing the model again, we get the results as

follow:

- SVM gave an accuracy of 99.42
- NB gave an accuracy of 97.33
- J48 gave an accuracy of 97.90

So, it is concluded that initially all three classifiers SVM, NB and J48 performed very well on given dataset but after adding prominent terms of fraud messages in dataset, it is observed that SVM outperformed all algorithms while Naïve Bayes and J48 dropped the accuracy. Fig. 3 expresses the results of this study where SVM (A), NB (A) and J48 (A) show the accuracy of the proposed model before adding extra normal messages having prominent terms of fraud messages and SVM (B), NB (B) and J48 (B) show the accuracy after adding prominent terms of fraud messages in dataset.

6. Conclusion and future work

Communication through SMS are increasing day by day. In the whole world, many fraudsters send spam/fraud messages to individuals to get some benefits from them. This research study focuses on some previous studies based on spam/fraud detection using supervised machine learning algorithms and introduces a new methodology to detect the fraud messages in Pakistan that are written using Roman Urdu. As there is no dataset available of Roman Urdu fraud messages so a manual dataset is created. WEKA tool is utilized to apply machine learning classifiers. The algorithms that are used in this study to detect fraud messages are SVM, Naïve Bayes and J48. 10-fold cross validation testing technique is used for each algorithm. It is observed that SVM gave the highest accuracy of 99.42

In this research study, we have used only two labels in our dataset which are 'normal messages' and 'fraud messages'. It can be further enhanced and multiple labels can be used like 'spam messages. Dataset can be further increased to get more accurate results. Different feature selection strategies can be applied to get more focused results. We have focused only on Roman Urdu fraud messages but multiple different local languages can also be utilized in dataset.

7. Acknowledgment

Mr. Ayaz's work is partly supported by his M.Phil. degree program in Institute of Mathematics and Computer Science University of Sindh Jamshoro.

References

[1] G. Mujtaba and M. Yasin, "Sms spam detection using simple message content features," *J. Basic Appl. Sci. Res*, vol. 4, no. 4, pp. 275–279, 2014.

[2] F. Pervaiz, R. S. Nawaz, M. U. Ramzan, M. Z. Usmani, S. Mare, K. Heimerl, F. Kamiran, R. Anderson, and L. Razaq, "An assessment of sms fraud in pakistan," in *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*, 2019, pp. 195–205.

[3] M. Group *et al.*, "Weka 3: Data mining software in java," *University of Waikato,[Online]. Available: http://www. cs. waikato. ac. nz/ml/weka/.[Accessed 10 Jan 2013]*.

[4] I. Wickramasinghe and H. Kalutarage, "Naive bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," *Soft Computing*, vol. 25, no. 3, pp. 2277–2293, 2021.

[5] J. Zhang, C. Li, Y. Yin, J. Zhang, and M. Grzegorzek, "Applications of artificial neural networks in microorganism image analysis: a comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer," *Artificial Intelligence Review*, vol. 56, no. 2, pp. 1013–1070, 2023.

[6] J.-S. Lee, "Auc4. 5: Auc-based c4. 5 decision tree algorithm for imbalanced data classification," *IEEE Access*, vol. 7, pp. 106 034–106 042, 2019.

[7] A. Karami and L. Zhou, "Improving static sms spam detection by using new content-based features," 2014.

[8] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[9] H. Tyralis, G. Papacharalampous, and A. Langousis, "A brief review of random forests for water scientists and practitioners and their recent history in water resources," *Water*, vol. 11, no. 5, p. 910, 2019.

[10] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of sms spam filtering: new collection and results," in *Proceedings of the 11th ACM symposium on Document engineering*, 2011, pp. 259–262.

[11] V. Bruni, M. L. Cardinali, and D. Vitulano, "A short review on minimum description length: An application to dimension reduction in pca," *Entropy*, vol. 24, no. 2, p. 269, 2022.

[12] D. A. Anggoro and N. D. Kurnia, "Comparison of accuracy level of support vector machine (svm) and k-nearest neighbors (knn) algorithms in predicting heart disease," *International Journal*, vol. 8, no. 5, pp. 1689–1694, 2020.

[13] K. Yılmaz and R. Tekin, "Comparison of discretization methods for classifier decision trees and decision rules on medical data sets," *Avrupa Bilim ve Teknoloji Dergisi*, no. 35, pp. 275–281, 2022.

[14] M. Khan, S. Ayub, M. Khan, M. Afaq, and F. Tila, "Machine learning for content-based detection of spam sms in local languages: A preliminary classification of romanized urdu messages," *Pakistan Journal of Science*, vol. 71, no. 4, p. 89, 2019.

[15] A. N. C. Abdul Rahman, I. H. Abdullah, I. S. Zainudin, S. Tiun, and A. Jaludin, "Domain-specific stop words in malaysian parliamentary debates 1959-2018." *GEMA Online Journal of Language Studies*, vol. 21, no. 2, 2021.

[16] K. Mehmood, H. Afzal, A. Majeed, and H. Latif, "Contributions to the study of bi-lingual roman urdu sms spam filtering," in *2015 National Software Engineering Conference (NSEC)*. IEEE, 2015, pp. 42–47.

[17] J. Su, H. Zhang, C. X. Ling, and S. Matwin, "Discriminative

parameter learning for bayesian networks," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1016–1023.

[18] E. A. Al-Zubaidi, F. Rabee, and A. H. Al-Sulttani, "Classification of large-scale datasets of landsat-8 satellite image based on liblinear library," *Al-Salam Journal for Engineering and Technology*, vol. 1, no. 2, pp. 9–17, 2022.

[19] H. Afzal and K. Mehmood, "Spam filtering of bi-lingual tweets using machine learning," in *2016 18th International conference on advanced communication technology (ICACT)*. IEEE, 2016, pp. 710–714.

[20] S. Nizamani, N. Memon, U. K. Wiil, and P. Karampelas, "Modeling suspicious email detection using enhanced feature selection," *arXiv preprint arXiv:1312.1971*, 2013.

[21] F. O. Adekunle, "A binary logistic regression model for prediction of feed conversion ratio of clarias gariepinus from feed composition data," *Marine Science and Technology Bulletin*, vol. 10, no. 2, pp. 134–141, 2021.

[22] F. Javed Mehedi Shamrat, R. Ranjan, K. M. Hasib, A. Yadav, and A. H. Siddique, "Performance evaluation among id3, c4. 5, and cart decision tree algorithm," in *Pervasive Computing and Social Networking: Proceedings of ICPCSN 2021*. Springer, 2022, pp. 127–142.

[23] S. Nizamani, N. Memon, M. Glasdam, and D. D. Nguyen, "Detection of fraudulent emails by employing advanced feature abundance," *Egyptian Informatics Journal*, vol. 15, no. 3, pp. 169–174, 2014.

[24] S. Nizamani, N. Memon, U. K. Wiil, and P. Karampelas, "Ccm: a text classification model by clustering," in *2011 International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 2011, pp. 461–467.

[25] S.-W. Kim and J.-M. Gil, "Research paper classification systems based on tf-idf and lda schemes," *Human-centric Computing and Information Sciences*, vol. 9, pp. 1–21, 2019.

[26] P. Karmani, A. A. Chandio, V. Karmani, I. A. Korejo, and M. S. Chandio, "Taxonomy on healthcare system based on machine learning," *International Journal of Computing and Digital Systems*, vol. 9, no. 6, pp. 1199–1212, 2020.

[27] P. Karmani, A. A. Chandio, I. A. Korejo, and M. S. Chandio, "A review of machine learning for healthcare informatics specifically tuberculosis disease diagnostics," in *Intelligent Technologies and Applications: First International Conference, INTAP 2018, Bahawalpur, Pakistan, October 23-25, 2018, Revised Selected Papers 1*. Springer, 2019, pp. 50–61.

Muhammad Ayaz is an M.Phil. scholar at the Institute of Mathematics and Computer Science, University of Sindh since 2019. Standing at the top of his class, he earned his BS (Hons) degree from the University of Sindh Campus in Mirpurkhas (2013-2016). He hopes to contribute to developments in the fields of machine learning and natural language processing. In addition, he is an accomplished computer programmer and has a deep interest in web development.

Dr. Sarwat Nizamani is currently working as Professor in Department of Computer Science at University of Sindh Mirpurkhas Campus. Dr. Nizamani received her M.Sc. (Hons.) degree in Computer Science from University of Sindh. She earned her Master of Science in Robotic System engineering in 2011 and Ph.D. in Information and Communication Technology in 2014 from University of Southern Denmark. Dr. Nizamani has published more than 35 research articles in Journals and Conferences of National and International repute. Most of her work is published by the prestigious publishers of the field such as, Elsevier, Springer and IEEE computer society. Moreover, she also published two book chapters published by the prestigious publisher Springer. Dr. Nizamani has also served as reviewer in numerous Journals and conferences of international repute.

Aftab Ahmed Chandio was born in Mehar city Dadu district, Sindh, Pakistan in 1983. He received the B.S. degree in computer science from the University of Sindh, Jamshoro, Pakistan, in 2007 and the Ph.D. doctoral engineering degree in computer application technology from Chinese Academy of Sciences, Beijing, China, in 2016.

Since 2007, he has been a permanent faculty member in Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Pakistan and now he is working as Associate Professor in the same institute. He is the author of more than 20 articles in journals and conferences, including CLUSCOMP, SUSCOM, FITEE, ZTE COMM, IEEE/ACM ISPA/TRUSTCOM, IEEE ICARCV, IEEE SCOReD, IEEE WORLD S4, IEEE WCICA, IOV in Springer's LNCS, INTAP in Springer's CCIS and IMECS in IAENG's LNECS. His research interests include resource management, job scheduling strategies, energy efficiency, and workload characterization for performance optimization of distributed systems such as cloud computing, and location-based services i.e., map-matching strategy for GPS trajectories. He is served as an invited reviewer in several journals and conferences, including IEEE Trans. on Cloud Computing, CLUSCOMP, Supercomputing, MONET, JPDC, IPCCC, IEEE CloudCom, IEEE ITSC and IEEE ICVES. Dr. Chandio was a recipient of the "Best Researcher Award 2019" of University of Sindh Jamshoro Pakistan, the "Best Paper Award" of 15th IEEE SCOReD 2017 Malaysia, the "Excellence Performance Award" as the Volunteer of IEEE/ACM CCGrid 2015 Shenzhen China and the "Dean Merit Scholarship" awards of Shenzhen Institutes of Advanced Studies Chinese Academy of Sciences for 2012 as well as for 2015.

Kirshan Kumar Luhana works as an Assistant Professor at the University of Sindh and is a researcher with a specific focus on Software Engineering, Extreme Programming, Artificial Intelligence, ICT4D (Information and Communication Technology for Development), and the integration of regional languages. He earned his PhD from TUG Austria in 2018.