Pakistan Journal of Science, 71 (4 Suppl.): 2019, Page 189-193
ISSN: 2411-0930

MACHINE LEARNING FOR CONTENT-BASED DETECTION OF SPAM SMS IN LOCAL LANGUAGES: A PRELIMINARY CLASSIFICATION OF ROMANIZED URDU MESSAGES

M.S. Khan[1, *], S.I. Ayub[1], M.A. Khan[1], M. Afaq[2] and F. Tila[3]

[1]University of Engineering & Technology, Mardan, Pakistan.
[2]Sarhad University of Science & IT, Peshawar, Pakistan.
[3]Bacha Khan University, Charsadda, Pakistan.
[*]Corresponding author's E-mail: sohail.khan@uetmardan.edu.pk

ABSTRACT: Recently, a drastic growth is being observed in mobile device services and its usage. Majority of the people rely on their cell phones for all types of communication. This increase in the use of mobile devices has led the network providers to provide more services on low cost. Short Message Service (SMS) is a widely used service among these services. SMS provides a very affordable medium of communication and hence is used by all sorts of businesses and organizations to stay in touch with their clients and many use the service for advertisement purposes. As a result, the average user of a mobile phone is receiving an increased number of SMS messages not relevant to her. The conventional spam SMS detection and avoidance techniques work better for English language but for SMS typed in local languages, such techniques are not so efficient. This study considers content-based spam SMS detection for local languages. As a preliminary investigation, a dataset of actual SMS typed in Romanized Urdu was collected for a number of cellphone users. The dataset was pre-processed and Support Vector Machine (SVM) was trained using the dataset for classification of spam and normal messages. The preliminary results are promising as the content-based classification for Romanized Urdu messages using SVM gives an accuracy of 96.8%.

Keywords: Machine Learning Classifiers, Short Message Service (SMS), Spam, Support Vector Machine (SVM).

INTRODUCTION

Short Message Service (SMS) is used to communicate in the form of short text messages. It is a widely used means of communication since the beginning of mobile phone. People communicate with their loved ones through SMSs more than on voice calls as voice calls are more expensive. The wide range of SMS use has led the network providers to give more and more services almost every day. In Pakistan SMS messages are used very frequently. According to a study carried by Pakistan Telecommunication Authority (PTA) in 2010, almost 151.6385 Billion text messages were exchanged during the year 2009 alone. The utility of SMS is not only for day-to-day communication but due to its affordability; SMS is used for advertisement purposes as well. Spam SMSs are annoying and irritating SMSs that are useless. These SMSs include advertisements, free services, promotions, awards, etc. People fall victim to these problems very easily. They give out their contact number to order food or participate in a lucky draw at a store but soon they start receiving advertisements and offers from these companies. These messages are annoying and useless for majority of the people. They want to shut them down but, unlike E-mail, they can't unsubscribe from receiving those messages. Apart from the advertisements from different companies, network providers send offers almost daily. Most of the offers they send are already availed by the person and most of them are not wanted by the person but they are sent repeatedly. Most of times irrelevant messages are repeatedly delivered to one's mobile phone at odd times of the day and night. Majority of the people fall victim to the frustration caused by these spam SMSs.

Similarly, Scam SMS are another consequence of this low-cost messaging service. Scam SMS is a message with a dishonest scheme. In the recent years' scammers in Pakistan have increased dramatically. More and more people are getting caught in fraudulent schemes. Fraudulent people use names of different TV shows that carryout lucky draws and trick people into their dishonest intentions. Mostly people with little or no education fall for their schemes losing valuable time and money as a result.

There have been several efforts to allow the user to avoid such frustrating messages either by blocking the senders number or by searching for specific keywords in an incoming message. Such applications and techniques are effective if the senders' mobile numbers are already in one's contact list or the message has been types in English language. When it comes to messages typed in local languages or Romanized local languages, the SMS blockers are ineffective. In this paper, an investigation in this regard is carried out. As Urdu language is the national language of Pakistan, Romanized Urdu messages are frequently exchanged among users as well as for advertisement messages sent by service providers and other businesses. As a preliminary study, messages typed in Romanized Urdu has been taken as a use-case to investigate whether machine learning techniques can be utilized for classification of SMS into spam and Non-spam messages.

To the best of our knowledge no dataset of Romanized Urdu SMS exists, thus a dataset has been created. For this purpose, an Android application was created and uploaded to Google play. Through this application SMSs are collected from real world mobile phone users. The dataset was then pre-processed and Support Vector Machine has been trained and tested using the dataset.

The rest of the paper contains the following sections. Section 2 provides an overview of related literature, highlighting any relevant studies. Section 3 presents the overall methodology of the research explaining how the dataset has been created and the process of training and testing the machine learning model has been illustrated. Section 4 discusses the results of the study and finally section 5 concludes the future with pointers to how the future work will be carried out for this study.

MATERIALS AND METHODS

Spam message recognition or classification for Romanized Urdu (or other local languages of Pakistan) has been the focus of much research work. These languages fall a little behind the world's developed languages. English, and other advanced languages have been thoroughly targeted in this regard and still progress in these languages is made every day. Some of the related work in this regard is given as follows.

(Rafique *et. al*., 2019) worked on the sentiment analysis of reviews made on movies in Romanized Urdu. They took data from different websites and preprocessed it by removing punctuation marks, stop words or numerical characters. Important features that are usually used in sentiment analysis were extracted from the preprocessed data. Three supervised machine learning techniques namely NB (Naive Bayes), LRSGD (Logistic Regression with Stochastic Gradient Descent), SVM (Support Vector Machine) were used to predict the mood of the Romanized Urdu text. The extracted features were fed to each of the models created from these techniques. In the end, the predicted results were compared with the expected outcomes to evaluate the accuracy of the proposed system. They got an accuracy of 87.22% for SVM and hence SVM stood out among the three techniques.

(Longzhen *et. al.,* 2009) has discussed some of the spam filtering approaches using the combination of both the K-Nearest-Neighbors (KNN) classification and Rough Sets to separate spam from legitimate messages. Rough set is used to remove redundant and un-necessary attribute from the data to improve classification accuracy. Reduction of features makes the decision making more efficient. The dataset they used have 550 spam messages while 200 normal text messages, a total of 750 message. The authors used precision, recall and F-measure as performance standards. By using k=12, they got precision of 91.35% while a recall of 84.50%.

Content based approaches used by (Zhang and Wang, 2006) are assumption based and they read the whole text of a message in order to classify the message as Spam or Normal. The study analyzed the SMS using features (tokens) and then decide it whether the SMS was legitimate or Spam. Content based filtering are of two kinds. Statistical based and rule based.

(Mathew and Issac, 2011) collected a corpus of 5000 SMS out of which 15% are spam messages. The authors used WEKA tool (Whitten *et al*., 2016) for experiments. As most of the algorithms cannot process text so they used 'StringtoWordVector' feature of WEKA which converts the text into word vectors. Best accuracy of 98.2% was achieved by Naïve Bayesian Multinomial, a variant of Bayesian algorithm. The authors preferred "Discriminative Multinomial Naïve Bayesian (DMNB) Text" over Naïve Bayesian because this algorithm returned an accuracy of 97.2% while giving the lowest false positive rate.

(Hidalgo *et al*., 2006) performed Bayesian filtering on SMS by collecting different SMS collections from different sources. Two different corpuses for two different languages, Spanish and English were collected. For English the authors combined SMS from John Stevenson Corpus (JSC), National University of Singapore NUS SMS Corpus and a UK based forum Grumble text and made a corpus of 1119 legitimate messages and 82 spam messages. Similarly, Yang performed feature selection on a corpus and selected attributes which scored more than 10 by using Information Gain (IG) (Yang and Pedersen, 1997; Yang, 1999). The dataset contained a total of 1203 messages where 82 were spam while 1119 were Normal messages. Naïve Bayes (Zhang and Wang, 2009), C4.5, PART (Aro *et. al.* 2019) and SVM were applied to the corpus for detection/classification of messages into spam and legitimate messages. The study found out that SVM was the best machine learning algorithm in their scenario as it produced a small number of false positives whereas performance of C 4.5 was the lowest.

It is evident from the above discussion that several efforts have been performed in order to classify English language and in some cases other European language text messages but not enough effort have been performed related to other local languages. Due to the rapid spread of ICT technologies and use of mobile services, it is high time to perform such investigations. The following section describes the methodology of the research in detail.

To carry out such a project one need to have real world data or the results will not be very accurate and meaningful. As there is no dataset of Text SMSs in Romanized Urdu available on the internet, actual data is collected from people. This data contains raw materials like repeated messages from advertisers and companies etc. Using some preprocessing it is reduced and made unique. The refined dataset was used to train for the classification of messages into spam and non-spam. The trained model was then tested using a subset of randomly selected messages from the dataset.

RESULTS AND DISCUSSION

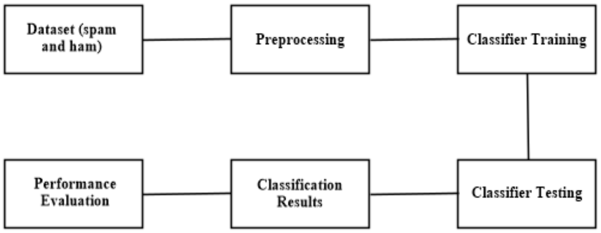

Figure 1: Basic Architecture

To collect data from people, an Android application was created and uploaded to Google play. This application has the following features:

- Shows user's all SMS messages in his/her phone.
- User selects the messages which he/she wants to contribute to the dataset.
- User has the option to mark a single SMS or whole conversation spam or non-spam.

Upon clicking the send button, the selected messages are sent to a Firebase Database.

Participants were told to download and use the application. Collected dataset contains 8289 messages. After preprocessing the dataset contained a total of 8107 messages including spam and non-spam messages. 7296 messages, which is approximately 90% of the contents of the dataset have been utilized for training while 10% or 811 messages have been used to test the model. The testing dataset contains 225 actual non-spam messages while 586 messages were spam messages. All those messages that are were not Romanized Urdu form were filtered out of the dataset. Table 1 provides a summary.

The dataset looks like this:

{
Label: spam/non-spam,
Text: actual text of the message.
}

Table 1: Dataset contents and data distribution.

Total Dataset	8289 Messages
After pre-processing	8107 Messages
Training	7296 Messages
Testing	811 Messages

Data cleaning comes before processing in many cases. It helps speed up the process. The data that is collect from users contains a lot of redundant or faulty content. The data is cleaned by removing the unnecessary content that do not help in identification or classification. For the dataset used in this study such as line breaks in the middle of text strings, messages that are not in Romanized form or repeated messages etc.

The dataset is converted to lowercase. Stop words are also removed in this step. Stop words are those which do not help in identifying if a message is a spam message or otherwise. The concept of stop words was first introduced by (Rafique *et al.*, 2019) and it has been utilized for preprocessing textual messages in almost all studies concerned with classification of SMS (Wilbur and Sirotkin, 1992) or messages from today's social media (Aro, *et al.*, 2019). Removing stop words speed up the computation process of the model. Particular to the current study, no list or dataset for stop words in Romanized Urdu exists so it was carefully created for this study by getting a list of stop words identified in Urdu script and transliterated them to Romanized Urdu using the resources at ijunoon. The list obtained in Romanized script was then further analyzed and corrected for any mis-transliteration. A sample of the list is given in Table 2.

Table 2: Sample of stop words in Romanized Urdu and Urdu Script.

Aa	آ
Aayi	آئی
Dekho	دیکھو
Rakha	رکھا
Mera	میرا

The dataset after cleaning shrinks down to 8107 messages. 90% of the messages has been used for training and 10% for testing. The study utilized Scikit library of the Python programming language. Support Vector Machine model from the Scikit library has been utilized for the study. The model classifies the messages based on how users labels a message. All the dataset is labeled spam or non-spam by participants of during the data collection process. The model reads the contents of message and checks its label and so it adjusts itself accordingly. One drawback of this dataset is that if a message for one person is ham but majority of the people labelled it as spam then the model classifies it as a spam message.

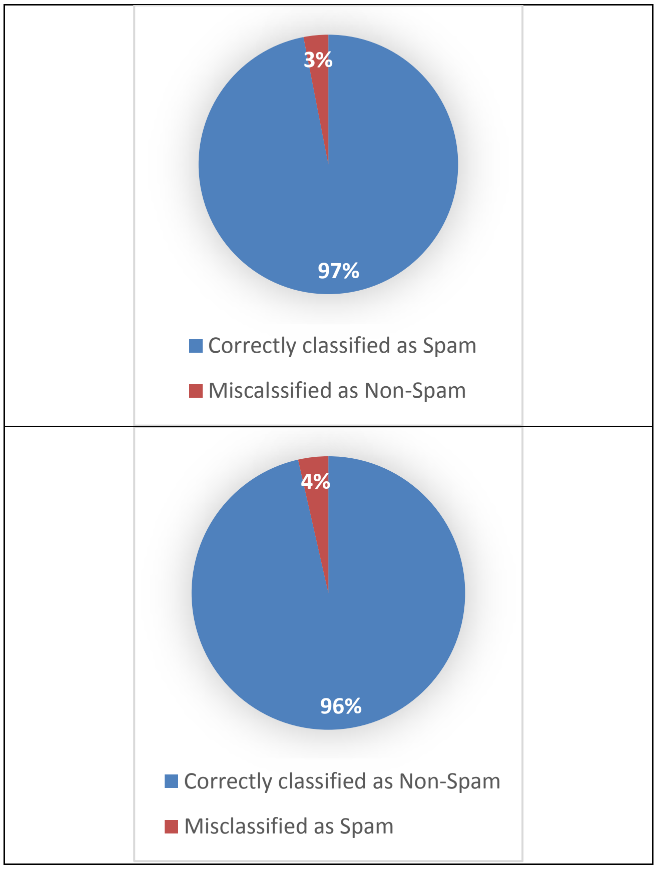

Figure 2: Correct and Misclassified messages

As mentioned earlier, out of all 8107 messages, approximately 10% i.e. 811 messages have been randomly selected for testing. Out of the selected 811 messages 586 are actually spam and 225 are actually non-spam messages. The trained model and the output of the model for each test-case was compared to the expected outcome i.e. spam or non-spam label. A confusion matrix was matrix for the model was obtained. The SVM model correctly predicts 568 of the spam messages as spam while 18 were misjudged as non-spam. For the non-spam messages in the testing set, 217 messages were correctly classified as non-spam while 8 non-spam messages were misjudged as spam. Figure 2 provides a clear illustration of the above results. Table 2 shows the accuracy and other performance measures of the model. The accuracy of the model is given by:

$$Accuracy = \frac{C}{P}$$

Where C represents the correct predictions and P represents the overall predictions by the model. In other words, accuracy is the ratio of the correctly labeled subjects to the whole pool of subjects. Precision is the fraction of relevant instances among the retrieved instances. Sensitivity is the proportion of actual positives that are correctly identified as such. Specificity is the proportion of actual negatives that are correctly identified as such.

Figure 3: Overall accuracy and performance measures for the trained SVM model.

Accuracy	96.8%
Precision	98.6%
Sensitivity	96.9%
Specificity	96.4%

Conclusion: SMS is widely used communication medium and it is being used for disseminating information and advertisements in Romanized local languages in order to reach a wider audience in countries where English language is not the first language. This paper focuses on investigating whether SMS typed in Romanized local languages e.g. Urdu, can be classified into spam and non-spam messages based on the content of messages. A dataset of spam and non-spam messages typed in Romanized Urdu language (National Language in Pakistan) were collected into a dataset and Support Vector Machine was utilized for the classification of the dataset. The preliminary results of the classification show promise with an accuracy rate of 96.8%. The dataset needs to be further refined and more training needs to be done using more attributes, for which the work is in progress. The future of the study includes, enhancing the current dataset, dataset creation for other local languages, training and testing via alternate machine learning models.

REFERENCES

Aro, T.O., F. Dada, A.O. Balogun and S.A. Oluwasogo (2019). Stop Words Removal on Textual Data Classification. *Development* 123, p.133.

Hidalgo, J., M.G. Bringas, G.C. Sánz, E.P. and F.C. García (2006). Content based SMS spam filtering. *in Proceedings of the 2006 ACM symposium on Document engineering* (pp. 107-114). ACM.

Longzhen, D., L. An and H. Longjun (2009). A New Spam Short Message Classification, *First International Workshop on Education Technology and Computer Science*, vol.2, no., pp.168,171.

Mathew, K. and B. Issac (2011). Intelligent spam classification for mobile text message, *in International Conference on Computer Science and Network Technology (ICCSNT*), vol.1, no., pp.101,105, 24-26.

Rafique, A., K. Malik, Z. Nawaz, F. Bukhari and A. Jalbani (2019). Sentiment Analysis for Romanized Urdu, *Mehran University Research Journal of Engineering & Technology*, doi: 10.22581/muet1982.1902.20.

Wilbur, W.J. and K. Sirotkin (1992). The automatic identification of stop words. *Journal of information science*, 18(1), pp.45-55.

Witten, I.H., E. Frank, M.A. Hall, and C.J. Pal (2016). Data Mining: Practical machine learning tools and techniques. *Morgan Kaufmann*.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization.," *in Information Retrieval*, 1(1/2): pp. 69-90.

Yang, Y. and J. Pedersen (1997). A comparative study on feature selection in text categorization, *in Proceedings of the 14th International Conference on Machine Learning*, In Icml (Vol. 97, No. 412-420, p. 35).

Zhang, H.Y. and W. Wang (2006). Lazy Associative Classification for Content-based Spam Detection, *Web Congress LA-Web*. Fourth Latin American , pp.154,161.

Zhang, H.Y. and W. Wang (2009). Application of Bayesian method to spam sms filtering, In *2009 International Conference on Information Engineering and Computer Science* (pp. 1-3). IEEE.