SMS SPAM DETECTION USING MACHINE LEARNING ALGORITHMS

SHAIK MOHAMMED IMRAN[1], NYTHANI HARSHITHA[2], MOHAMMAD YASMEEN[3], MITTAPELLY RUTHIKA[4]

ASSISTANT PROFESSOR[1], UG SCHOLAR[2,3&4]

DEPARTMENT OF CSE, MALLA REDDY ENGINEERING COLLEGE FOR WOMEN,MAISAMMAGUDA, DHULAPALLY KOMPALLY, MEDCHAL RD, M, SECUNDERABAD, TELANGANA 500100

ABSTRACT

With the prosperity of the Short Message Service (SMS), the increasing number of spam messages has become a serious problem. The need to block spam messages requires us to develop new SMS spam detection technologies. The Transformer, an attentionbased sequence to sequence model, has achieved excellent results in multiple different tasks recently. In this thesis, we propose a modified Transformer model for SMS spam messages detection. The evaluation of our proposed modified spam Transformer is performed on SMS Spam Collection v.1 dataset and UtkMl's Twitter Spam Detection Competition dataset, with the benchmark of multiple established classifiers such as Logistic Regression, Na¨ıve Bayes, Random Forests, Support Vector Machine, and Long Short-Term Memory. In comparison to all other candidates, our experiments show that the proposed modified spam Transformer achieves the best results in terms of almost all selected performance criteria.

I.INTRODUCTION

The Short Message Service (SMS) has been widely used as a communication tool over the past few decades with the popularity of mobile phones and mobile network growth. According to Statista's report in 2020, the number of mobile users worldwide reaches 6.95 billion with the forecasts indicating an increase to 7.1 billion by 2021, [56]. A survey in 2019 shows that more than 76% of mobile users in UK send mobile messages on a daily basis [70]. Mobile devices and SMS serverces are becoming necessities of people's daily life, The Guardian reports that people pick up their smartphones 58 times a day on average [47] and Finance Online reveals that more than 90% of people prefer to receive messages over calls [19]. With the development of mobile devices and various mobile applications, billions of people are connected and millions of businesses and marketers are benefited from trillions of SMS text messages worldwide. However, SMS users are also suffering from SMS spam. SMS spam, also known as drunk messages, refers to any irrelevant messages delivered using mobile networks [66]. Statista conducted a survey in the United States from March 20 to March 24 last year, and the survey points out that the number of spam messages growing exponentially over the past years with an average number of 14.7 spam text messages per month [72]. The Government of Canada defines spam as unsolicited text messages including malware, spyware, and false or misleading representations [10]. There are a few characteristics of spam text messages summarized by the Federal Trade Commission. In general, spam text messages aim to get your personal information by sending some fake text messages and end up gain financial benefits [75]. There are several reasons that lead to the popularity of spam messages. First and foremost, there is a large number of users who use mobile phones in the world. According to [71], there were 8,304 million mobile subscriptions in 2019. A large amount of mobile phone users makes the potential victims of the spam messages attack also high. Secondly, the cost of sending out spam messages is low compared to the potential gains, which could be taken advantage of by the spam attacker. Last but not least, the capability of the spam classifier on most mobile phones is relatively weak due to the shortage of computational and battery resources on portable devices, which limits them from identifying the spam messages correctly and efficiently. Machine learning is one of the most popular topics in the last few decades. The applications of machine learning techniques have been greatly changing our lives. Spam detection is a relatively mature research topic in the field of machine learning, and there are a great number of machine learning based approaches proposed. Most of the machine learning based classifiers were dependent on the handcrafted features extracted from the training data [38]. As a class of machine learning techniques, deep learning has been developing rapidly recently thanks to the surprising growth of computational resources in the last few decades. Nowadays, deep learning based applications play a significant part in our society, making our lives much easier in many aspects. As one of the most effective and widely used deep learning architectures, Recurrent Neural Network (RNN), as well as its variants such as Long Short-Term Memory (LSTM), were applied to spam detection and proved to be extremely effective during the last few years. Although the current SMS spam detection

approaches have been proved to be effective in multiple experiments performed by different researchers, most of them are still merely focus on one single small unbalanced dataset. In reality, a lot of the existing works are based on the SMS Spam Collection v.1 [1] dataset, which is a fabulous dataset with high-quality real messages but is insufficient in terms of the number of spam messages. Therefore, we not only intend to propose a new approach for SMS spam detection that improves the performance compared to the current works. More importantly, we expect to test and evaluate our methodology on a larger dataset with more balanced data.

II.LITERATURE REVIEW

There are several different machine learning based classification applications proposed in the last few decades [6], [7] [8], [9]. In the field of SMS spam detection, a great number of these approaches are based on traditional machine learning techniques, such as Logistic Regression (LR), Random Forest (RF) [10], Support Vector Machine (SVM) [11], Naïve Bayes (NB), and Decision Trees (DT). Recently, with the prosperity of the deep learning techniques, an increasing number of methods have been introduced to address the SMS spam problem using deep learning based solutions such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM), which is a successful variant of RNN

In [12], Gupta et al. compared the performance of 8 different classifiers including SVM, NB, DT, LR, RF, AdaBoost, Neural Network, and CNN. The experimental tests on the SMS Spam Collection v.1 [13] dataset that was conducted by the authors shows that the CNN and Neural Network are better compared to other machine learning classifiers, and the CNN and Neural Network achieved an accuracy of 98.25% and 98.00%, respectively

In [14], Jain et al. proposed a method to apply rule-based models on the SMS spam detection problem. The authors extracted 9 rules and implemented Decision Tree (DT), RIPPER [15], and PRISM [16] to identify the spam messages. According to the experimental results from the authors, the RIPPER outperformed the PRISM and the DT, yielding a 99.01% True Negative Rate (TNR) and a 92.82% True Positive Rate (TPR)

In [1], Roy et al. aimed to adapt the CNN and LSTM to the SMS spam messages detection problem. The authors evaluated the performance of CNN and LSTM by comparing them with Naïve Bayes (NB), Random Forest (RF), Gradient Boosting (GB) [17], Logistic Regression (LR), and Stochastic Gradient Descent (SGD) [18]. The experiments that were conducted by the authors showed that the CNN and LSTM perform significantly better than the tested traditional machine learning approaches when it comes to SMS spam detection

In [2], the authors proposed the Semantic Long Short-Term Memory (SLSTM), a variant of LSTM with an additional semantic layer. The authors employed the Word2vec [19], the WordNet [20], and the ConceptNet [21] as the semantic layer, and combined the semantic layer with the LSTM to train an SMS spam detection model. The experimental evaluation that was conducted by the authors claimed that the SLSTM achieved an accuracy of 99% on the SMS Spam Collection v.1 dataset.

In [22], Ghourabi et al. proposed the CNN-LSTM model that consists of a CNN layer and an LSTM layer in order to identify SMS spam messages in English and Arabic. The authors evaluated the CNN-LSTM by comparing it with the CNN, LSTM, and 9 traditional machine learning solutions. The experimental tests that were conducted by the authors showed that the CNN-LSTM solution performed better than other approaches and yield an accuracy of 98.3% and an F1-Score of 0.914

III.PROPOSED SYSTEM

The Transformer [3] is an attention-based sequence-to sequence model that was originally designated for translation task, and it achieved great success in English-German and English-French translation. Moreover, there are multiple improved Transformer-based models such as GPT-3 [4] and BERT [5] proposed recently to address different Natural Language Process (NLP) problems. The accomplishments of the Transformer and its successors have proved how powerful and promising they are. In this paper, we aim to explore whether it is possible to adapt the Transformer model to the SMS spam detection problem. Therefore, we propose a modified model based on the vanilla Transformer to identify SMS spam messages. Additionally, we analyze and compare the performance of SMS spam detection between traditional m achine learning classifiers, an LSTM deep learning solution,and our proposed spam Transformer model.

IV.SYSTEM ARCHITECTURE

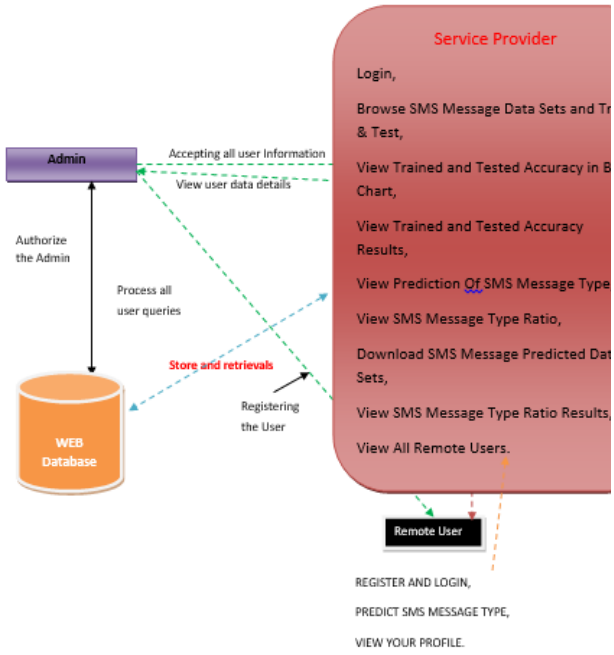

Fig 1: System Architecture

As We Discussed In The Fig 1 The Things Are Classified And Described Below

- Service Provider
- View And Authorize User
- Remote User

V.DESCRITPION

SERVICE PROVIDER

- In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as
- Browse SMS Message Data Sets and Train & Test, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results,
- View Prediction Of SMS Message Type, View SMS Message Type Ratio,
- Download SMS Message Predicted Data Sets, View SMS Message Type Ratio Results, View All Remote Users.

VIEW AND AUTHORIZE USERS

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

REMOTE USER

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like PREDICT SMS MESSAGE TYPE,VIEW YOUR PROFILE.

VI.ALGORITHMS

Decision tree classifiers

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class
Step 2. Otherwise, let T be some test with possible outcomes O1, O On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2,… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

K-Nearest Neighbors (KNN)

- Simple, but a very powerful classification algorithm
- Classifies based on a similarity measure
- Non-parametric
- Lazy learning
- Does not "learn" until the test example is given
- Whenever we have a new data to classify, we find its K-nearest neighbors from the training data

Example

- Training dataset consists of k-closest examples in feature space

- Feature space means, space with categorization variables (non-metric variables)
- Learning based on instances, and thus also works lazily because instance close to the input vector for test or prediction may take time to occur in the training dataset

Logistic regression Classifiers

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

Naïve Bayes

The naive bayes approach is a supervised learning method which is based on a simplistic hypothesis: it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature .

Yet, despite this, it appears robust and efficient. Its performance is comparable to other supervised learning techniques. Various reasons have been advanced in the literature. In this tutorial, we highlight an explanation based on the representation bias. The naive bayes classifier is a linear classifier, as well as linear discriminant analysis, logistic regression or linear SVM (support vector machine). The difference lies on the method of estimating the parameters of the classifier (the learning bias).

While the Naive Bayes classifier is widely used in the research world, it is not widespread among practitioners which want to obtain usable results. On the one hand, the researchers found especially it is very easy to program and implement it, its parameters are easy to estimate, learning is very fast even on very large databases, its accuracy is reasonably good in comparison to the other approaches. On the other hand, the final users do not obtain a model easy to interpret and deploy, they does not understand the interest of such a technique.

Thus, we introduce in a new presentation of the results of the learning process. The classifier is easier to understand, and its deployment is also made easier. In the first part of this tutorial, we present some theoretical aspects of the naive bayes classifier. Then, we implement the approach on a dataset with Tanagra. We compare the obtained results (the parameters of the model) to those obtained with other linear approaches such as the logistic regression, the linear discriminant analysis and the linear SVM. We note that the results are highly consistent. This largely explains the good performance of the method in comparison to others. In the second part, we use various tools on the same dataset (Weka 3.6.0, R 2.9.2, Knime 2.1.1, Orange 2.0b and RapidMiner 4.6.0). We try above all to understand the obtained results.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

The first algorithm for random decision forests was created in 1995 by Tin Kam Ho[1] using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.).The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho[1] and later independently by Amit and Geman[13] in order to construct a collection of decision trees with controlled variance.

Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

SVM

In classification tasks a discriminant machine learning technique aims at finding, based on an *independent and identically distributed* (*iid*) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point *x* and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to *genetic algorithms* (*GAs*) or *perceptrons*, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

VII.RESULTS

Fig 2: Home Page

Fig 3: Login page

Fig 4: User Registration page

Fig 5: View Remote User

Fig 6: Pi Chart

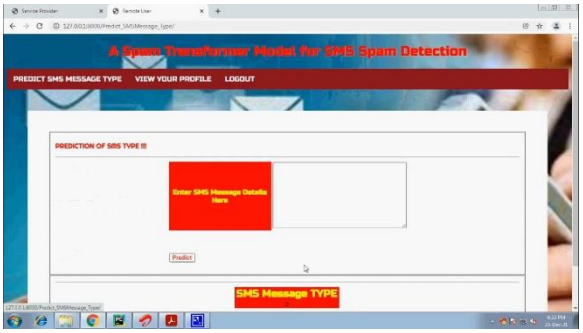

Fig 7: Predict SMS Message Type

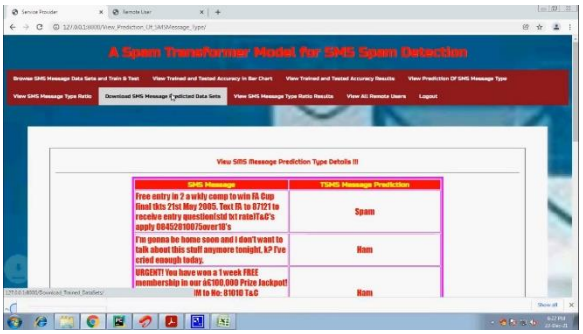

Fig 8: Displaying Type of SMS

CONCLUSION

In This Paper,We proposed a modified Transformer model that aims to identify SMS spam based on the Transformer [77] model. We evaluated our spam Transformer model by comparing it with several other SMS spam detection approaches on the SMS Spam Collection v.1 dataset [1] and UtkMl's Twitter dataset [76]. The experimental results show that, compared to Logistic Regression, K-Nearest Neighbor, Na¨ıve Bayes, Random Forests, Support Vector Machine, Long Short-Term Memory, and CNN-LSTM [26], our proposed spam Transformer model performs better on both datasets. On the SMS Spam Collection v.1 dataset [1], our spam Transformer has a better performance in terms of accuracy, recall, and F1-Score compared to other classifiers. Specifically, our modified spam Transformer approach accomplished an exceeding result on F1-Score. Additionally, on the UtkMl's Twitter dataset [76], the experimental results acquired from our modified spam Transformer model demonstrate its improved performance on all four aspects in comparison to other alternative approaches mentioned in this thesis. Concretely, our spam Transformer does exceptionally well on recall, which contributes to a distinct F1-Score

REFERENCES

[1] P. K. Roy, J. P. Singh, and S. Banerjee, ''Deep learning to filter SMS spam,'' Future Gener. Comput. Syst., vol. 102, pp. 524–533, Jan. 2020.

[2] G. Jain, M. Sharma, and B. Agarwal, ''Optimizing semantic LSTM for spam detection,'' Int. J. Inf. Technol., vol. 11, no. 2, pp. 239–250, Jun. 2019.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5999–6009.

[4] T. B. Brown et al., ''Language models are few-shot learners,'' 2020, arXiv:2005.14165. [Online]. Available: http://arxiv.org/abs/2005.14165

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., vol. 1, Jun. 2019, pp. 4171–4186.

[6] G. Sonowal and K. S. Kuppusamy, ''SmiDCA: An anti-Smishing model with machine learning approach,'' Comput. J., vol. 61, no. 8, pp. 1143–1157, Aug. 2018.

[7] J. W. Joo, S. Y. Moon, S. Singh, and J. H. Park, ''S-detector: An enhanced security model for detecting Smishing attack for mobile computing,'' Telecommun. Syst., vol. 66, no. 1, pp. 29–38, Sep. 2017.

[8] S. Mishra and D. Soni, ''Smishing detector: A security model to detect Smishing through SMS content analysis and URL behavior analysis,'' Future Gener. Comput. Syst., vol. 108, pp. 803–815, Jul. 2020.

[9] C. Li, L. Hou, B. Y. Sharma, H. Li, C. Chen, Y. Li, X. Zhao, H. Huang, Z. Cai, and H. Chen, ''Developing a new intelligent system for the diagnosis of tuberculous pleural effusion,'' Comput. Methods Programs Biomed., vol. 153, pp. 211–225, Jan. 2018.

[10] T. K. Ho, ''Random decision forests,'' in Proc. Int. Conf. Document Anal. Recognit. (ICDAR), vol. 1, 1995, pp. 278–282.

[11] C. Cortes and V. Vapnik, ''Support-vector networks,'' Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[12] M. Gupta, A. Bakliwal, S. Agarwal, and P. Mehndiratta, ''A comparative study of spam SMS detection using machine learning

classifiers,'' in Proc. 11th Int. Conf. Contemp. Comput. (IC3), Aug. 2018, pp. 1–7.

[13] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, ''Contributions to the study of SMS spam filtering: New collection and results,'' in Proc. 11th ACM Symp. Document Eng., Sep. 2011, pp. 259–262.

[14] A. K. Jain and B. B. Gupta, ''Rule-based framework for detection of Smishing messages in mobile environment,'' Procedia Comput. Sci., vol. 125, pp. 617–623, 2018.

[15] W. W. Cohen, ''Fast effective rule induction,'' in Machine Learning Proceedings, 1995, pp. 115–123.