Twitter Spam Detection using Pre-trained Model

Ankur Gupta, Yogendra P.S. Maravi, Nishchol Mishra

***Abstract:** In the age of technology social media platform is becoming a great companion for expressing the thoughts, information, and opinion. It became the powerful tool for every person who wants to expand their networks of people beyond the physical boundation. We are living at that age where various categories of social media platform available according to work needed, it may be Facebook, LinkedIn, WhatsApp or Twitter. We are focusing our work on Twitter, It is also known as microblogging site which provides service to express the opinion in limited words. As the popularity of twitter is growing day by day users are joining the platform very fast, as it happens another side many spammers are also taking undue advantages of this platform, for any social media platform it is very important to maintain the secure, safer and trustworthy environment for their legitimate users. Twitter spams are more harmful than e-mail spam because of their higher clickthrough rate, as in the social network if someone trusted some spam a genuine post than it is higher chance that the persons in the network might also trust on that spam post and may click on it. There are plenty of methods available to handle the task of twitter spam detection problem, we are solving this problem of twitter spam at tweet level.Pre-trained models are some breakthrough in the journey of machine learning and natural language processing after their advancement they are of great help. Here we are using Bidirectional Encoder Representation from Transformer (BERT) model to solve the problem as our task is to solve the problem of imbalance dataset as well as the multilingual dataset, BERT makes a clear distinction in this type of task, the main advantage of this type's model is that we don't have to collect millions of data for better performance of the machine learning model.*

***Keywords:** Twitter, BERT, spam detection, Pre-trained*

I. INTRODUCTION

We are living in the age of technology where social media plays crucial roles in sharing the information, thoughts, personal opinion, critics . Twitter is one of the popular social media platform also known as the microblogging site gives the freedom to express our thoughts within 280 characters. Spammers also try to pollute the platform by sending various tweet having malicious links, irrelevant tweets, attractive tweets to increase the followings and other purposes.

It is known that spam tweet is more harmful than a spam e-mail because of higher clickthrough rate of a tweet[9]. Many spam detection methods are based on the blocking spammer, it is also worth noting that if we block the spammer than they may create new account and starts posting spam tweets so it is not a good idea to only block the spammer we also need some solution which can works directly on the tweets whether it is tweeted by spammer or genuine users.

It is also a question that if we have blacklisting services for analyzing the malicious URL then why we are using the learning techniques. Blacklisting services have their own associated disadvantages e.g. detection of a URL whether it is malicious or not takes a long time, and till then user may click the link and that link may spread to the networks of users thus using blacklisting services are also not more effective. In real scenario spam to non-spam tweet ratio is very low so it is the challenge to make methods that can perform well on the imbalance dataset. Due to the imbalance dataset problem various model biased towards majority class or with higher margin biased towards minority class it is tough to achieve balanced metric. Another problem which is identified here is that tweet can be in various language rather than English so handling these problem with higher efficiency is challenging task too.

II. RELATED WORK

Many researchers solved the problem of spam in twitter platform at different levels, Wu et al.[4] worked on this problem using word embedding specifically word2vec[14] using word2vec for feature extraction in the form of vector and also used perceptron for the middle layer they solved the problem up to very much extent. Main benchmark happen in the field of twitter spam detection at tweet level is when word embedding like word2vec developed it helped the research work by defining the various vectors of the word actually word embedding helped a lots because it describe the meaning of the words in a sentence where another methods describes like in one hot encoding a simple atomic word if we write 'he is playing football' in terms of global word embedding(word2vec,glove) the relatedness of playing and of football is closer to one which is also sometimes called as cosine similarity but there is no similarity we can find with applying one-hot encoding. Sedhai et al[5] have given an insight into the problem where we have fewer datasets, where some dataset are not labelled and some are labelled they first trained the model with labelled ones and by using them tried to label unlabeled data in semi-supervise learning way. They used their own Hspam14dataset [13]. Using the semi-supervised way of learning is a very complex task.

Manuscript published on November 30, 2019.
* Correspondence Author

Ankur Gupta*, School of Information Technology, RGPV, Bhopal, India. Email: ankurgpt70@gmail.com
Yogendra P.S. Maravi, School of Information Technology, RGPV Bhopal, India. Email:Yogendra.rgpv@gmail.com
Nishchol Mishra, School of Information Technology, RGPV Bhopal, India. Email: nishchol@rgtu.net

Some researchers also used a hybrid approach where they applied the method to detect spam profiles on twitter, all the work which authors have done is in real-time based detection system which is the need of the system .

Authors in [7] tell about what type of methods available for the spam detection in twitter at what level, It describes the scenario of research at the tweet level, user level, at tweet level generally focus is on the tweet content, Bag of word is important method on that.

V.Vishvarupe et al.[2] used Google safe Browsing API in their architecture, they considered some useful feature sets like unsolicited mentions, links to malicious URLs, duplicate domain names etc,they proposed a system where they looked for a particular hashtags.

Ameen et al.[6] used deep learning method with binary classifier, they used the word embeddings with MLP(Multi-Layer Perceptron) and compared with other classifiers like Naïve Bayes, Random Forest, Decision Tree .Their approach was extracting feature by Word2vec and then applying learning algorithm ,their approach gave some good results in terms of recall, precision and f-measure .

Medisety et al[1] describes the tweet level spam detection task .They have used two dataset first is imbalance and another one is totally balanced dataset. They analyzed both word embedding with CNN(Convolutional Neural Networks) and one method which is totally based on feature analysis and thus making the ensemble stack of deep learning models with a feature analysis machine learning model, thus they were able to produce a good result of 89% in f-measure but not so good what today a pre-trained model can produce.

It is also worth noting that [11] authors described the various method like lstm and cnn to solve this problem and shown a good result from that.

There is also various methods available to detect spammers if we detect spammer we can block them but some time spammer, third party app takes authorization of legitimate user and on behalf of them tweet with malicious post so it is more valuable to work on the tweet level spam detection .As natural language processing area is developing day by day researcher community is working on the solving the language related problem in more intelligent way as a human can solve, in that process more language models are being developed and tested BERT ,Open AI GPT,ELMo etc are the examples of intelligent language model.

III. METHODOLOGY

Pre-trained models are the models which are generally trained on large dataset we have to only fine-tune the particular dataset and they learn from that small dataset.

As we are working with twitter spam so we generally have more tweets in non-spam category and very less tweets are spam category-trained model can also be treated as an example of transfer learning where a model learn from different dataset and that learning is applied to another dataset. For natural language processing task it is most usable technique to get more efficient than any other general purpose neural network like CNN,RNN(Recurrent Neural Network),LSTM(Long Short term Memory).

Various type of pre-trained model are being used by various industrial organisation because it is totally a language model which are designed for the natural processing task. For our work we are using the BERT[12] model which is known as Bidirectional Encoder Representation from Transformer. This model works bidirectionally where other model works unidirectionally. It has twelve layers and twenty four layers version. There are various version of BERT is available like bert base case ,bert base uncased ,bert large cased ,multilingual .Here we are using BERT multilingual cased version. BERT multilingual model is trained on various languages article .Our dataset contains some another language's tweet other than the English so it is better to use BERT with multilingual version for better result.

It has achieved benchmark performance in various linguistic task .Main benefit of using these type of model is that we have to only fine-tune our datasets. We don't have to train the model from scratch which is more crucial .BERT could be fine-tune with just one extra output layer .It is simple and effective. It uses self-attention mechanism. Advantage of using this model is that it has been trained on millions of words so it has the better understanding of every word in the text and we also have less amount of data so using this model we can compensate with data and also we can get more better results. We just give the task specific inputs BERT fine tune all the parameter end-to-end .We generally need high processing unit for the fine-tuning the model like BERT.

This model is already trained on eight hundred millions of word and twenty five hundreds millions of word from book corpus and Wikipedia respectively.

As it is masked language model here every input having some mask to predict the next word. It is trained like 'He is playing football' then mask will be He and training data contains _is playing football BERT have to predict word He thus making combination of various sentences it is pretrained on millions of sentences from Wikipedia and book corpus.In our task it is basically classification task we have to train with mask [CLS] where BERT has to predict either 1 for spam or 0 for non-spam tweet .We are also using SoftMax function on top of the model for classification task. Because BERT is made of 12 transformer head which works totally on attention mechanism so it is very good model for language task like tweet spam classification.

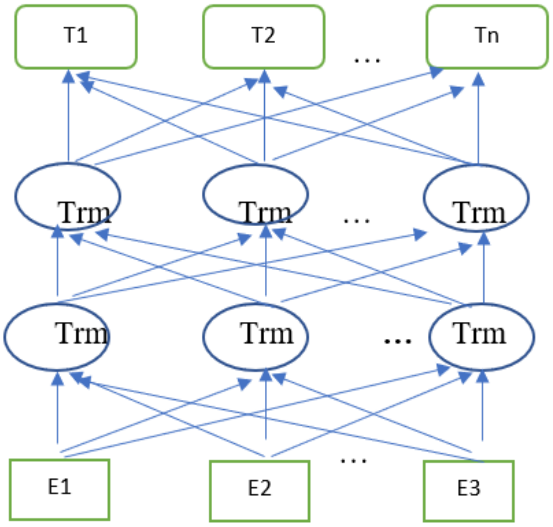

Fig.1.Biderectional Encoder Representation from Transformer(BERT)

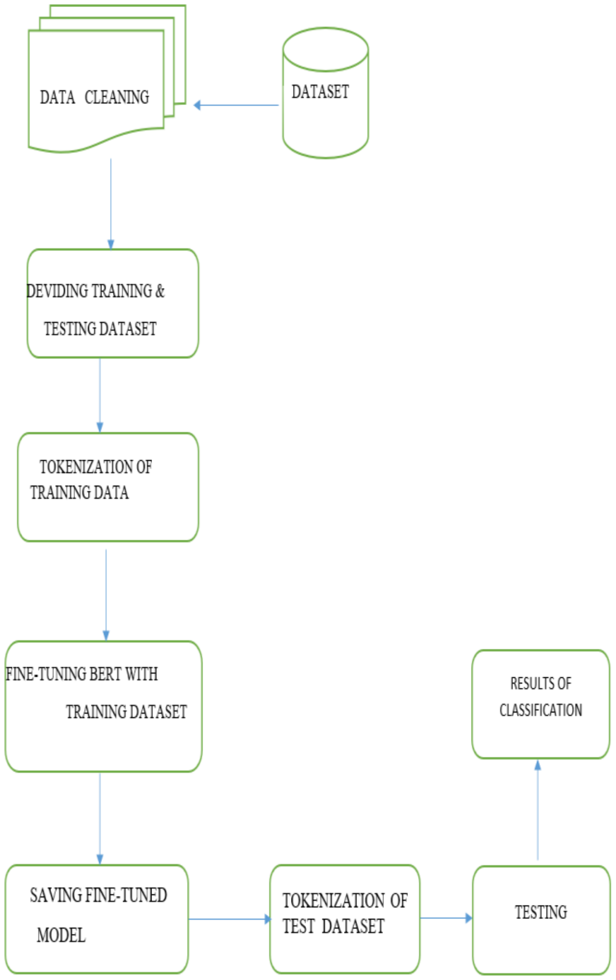

Fig. 2. Mechanism to be followed for detection task .

IV. EXPERIMENT AND RESULTS

We performed our experiments with 64 GB RAM, Intel xeon processor of 2.7 GHZ clock frequency.

Here we are going to evaluate our proposed BERT method for twitter spam detection, before discussing the experimental results, we describe about the dataset and evaluation metrics.

A. *Dataset:*

For any machine learning task dataset plays an important role, for our dataset we are taking a benchmark 1KS10KNS[10] dataset as this dataset is basically contains tweets of one thousand spammer's tweet and ten thousands non spammer's tweet so it can be termed as imbalance dataset. This dataset contains tweet in English as well as in Portuguese language previous researchers only focused on non-language model like CNN,RNN, LSTM where they use tokenization from only English language ,in this paper we took default tokenization of BERT multilingual model which is capabale of tokenizing multilingual words in one sentence because a single tweet can have words of various language also. We took training and testing data of 68 % for training and remaining for testing.

B. *Metrics For Evaluation:*

TABLE I
CONFUSION MATRIX FOR CLASSIFICATION

	SPAM	NON-SPAM
SPAM	TP	FN
NON-SPAM	FP	TN

Evaluation of model is the core part of model creation we are taking true positive(TP) to the number of spam which are actually spam and model predicted them spam, true negative (TN) to the numbers of spam which are actually non-spam tweet instance and predicted as same ,false positive(FP) for non-spam tweet which is classified as spam by model and false negative(FN) is the number of spam tweet which were predicted as non-spam by model .Basically these are the four parameter of confusion metrics which is generally used to evaluate a model.

We took important parameter of evaluation as accuracy ,precision,recall,f-measure. Here it is important to note that for any imbalance dataset model works with bias toward majority and minority as well,if they bias towards minority than values of recall will be higher like in case of SVM generally pick the minority class and with higher margin gives the recall values higher and precision values very low but in our case in spite of higher recall values precision is also higher thus producing a good result.

F-measure which is the indicator of harmonic mean of the precision and recall .BERT is very good at finding minor class as well as major class previous attempt by some researcher show good f-measure but using stack of CNN and feature based model thus ensembling various model but here we are using only one model for the task and producing better result.

TABLE II
RESULT IN CONFUSION MATRIX

	SPAM	NON-SPAM
SPAM	303	27
NON-SPAM	18	3181

TABLE III
RESULT COMPARISON

Methods	Accuracy	Precision	Recall	F-measure
CNN-Ensemble	0.99	0.92	0.86	0.89
Proposed	0.99	0.94	0.92	0.93

We took all the experiment using python language ,for data cleaning task we used pandas library which is used as the advanced version of excel to handle the data as well as to represent the dataset, for numerical analysis we took the help of numpy library of the python which used as one of the most numerical library in machine learning community and deep learning part of implementation of BERT we too the help of pytorch library of python which is most usable library in academic by python community for deep learning task .BERT is implemented in python and included in the pytorch library of standard python, We just have to call the library and fine-tune our dataset of our task by setting various hyperparameter like we set droupout-0.1,sequence length 130,batch size 24.

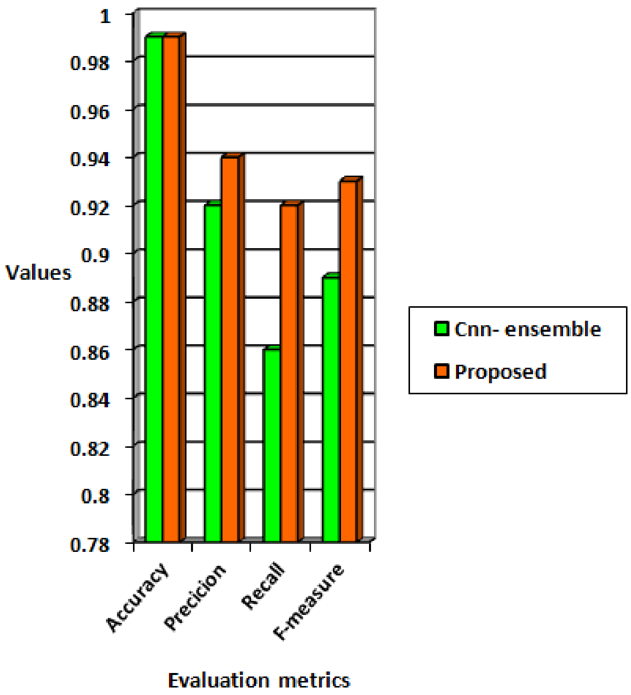

Fig.3.Graph showing result of classification

V. CONCLUSION

In this paper we have proposed new approach in twitter's spam problem at tweet level .Our approach used the model of transfer learning mechanism ,by the help of pretrained model we can use the model which were trained on some other dataset but we can fine tune them with required dataset. After applying the BERT multilingual base method which has twelve layers gives us good metrics of evaluation. As our dataset contained tweet in another language ,in spite of that our approach worked well to classify the spam and non-spam tweet. We think it will be better approach to solve the problem where we have imbalance dataset in various language. The proposed method can be further extended to explore some new dataset in Indian language like Hindi, Tamil etc so that it can be helpful in making the twitter platform more trustworthy social media platform.

REFRENCES

1. Sreekanth Madisetty, M. S. Desarkar.(2018) "A neural network based ensemble approach for spam detection in twitter" IEEE Transaction on Computational Social Systems. Volume(5).pages 973 – 984
2. Varad Vishwarupe, Mangesh Bedekar, Milind Pande and Anil Hiwale.(2017) "Intelligent Twitter Spam Detection: A Hybrid Approach" Springer Smart Trends in Systems, Security and Sustainability,pages 189–197.
3. Chaoling Li, and Shigang Liu.(2017) "A comparative study of the class imbalance problem in Twitter spam detection" wiley Concurrency and Computation: Practice and Experience.
4. Tingmin Wu, Shigang Liu, Jun Zhang, & Yang Xiang.(2017) "Twitter spam detection based on deep learning"In ACSW Multiconference.
5. Surendra Sedhai & Aixin Sun.(2017) "Semi-supervised spam detection in Twitter stream"IEEE Transaction on Computational Social System. volume (5).pages169–175.
6. Aso K Ameen., & B.Kaya.(2018) "Spam detection in online social networks by deep learning"IEEE, Int. Con. on A.I and Data Processing
7. T. Wu, S. Wen, W. Zhou&Y. Xiang.(2017) "Twitter spam detection: Survey of new approaches and comparative study"Springer. Computer Security., volume (76).pages 265–284.
8. Kurt Thomas ,Chris Grier ,Justin Ma ,Vern Paxson ,Dawn Song.(2011) "Design and evaluation of a real-time URL spam filtering service" In IEEE Symp. S P.pages 447–462.
9. Chris Grier, Kurt Thomas, Vern Paxson, and M. Zhang.(2010) "@spam: The underground on 140 characters or less" In 17th ACM Conf. of C.C. S.pages 27–37.
10. Chao Yang, Robert Harkreader, Jialong Zhang, Seingwon Shin, & Gufei Gu.(2012) "Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on Twitter" In 21st International Conf. of W WW.pages 71–80.
11. Gauri Jain, Manish Sharma,Basant Agarwal.(2019) "Spam detection in social media using convolutional and long short term memory neural network"In Annals of Maths & A.I.Volume (85), Issue 1.pages 21–44.
12. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova .(2018) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding."In the arxiv.com
13. Surendra Sedhai & A. Sun.(2015) "HSpam14: A collection of 14 million tweets for hashtag-oriented spam research"In 38th International ACM SIGIR Conf.pages 223–232.
14. Q. Le and T. Mikolov.(2014) "Distributed representations of sentences & documents" In 31st Int. Conf. M. L.. pages 1188–1196.
15. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean.(2013) "Distributed representations of words and phrases and their compositionality" In 26th Int. Conf. of N.I.P.S., pages 3111–3119.

AUTHORS PROFILE

Ankur Gupta is currently pursuing M.Tech. degree in computer technology and application from school of information and technology, RGPV, Bhopal, India. He has completed his B.E. from RGPV, Bhopal, India in year 2014.His current research interests are Machine Learning, Natural Language Processing, Quantum Computing.

Yogendra P.S Maravi is currently an Assistant Professor at School of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India. He has completed his B.E.in Information Technology from UIT, RGPV Bhopal and M.Tech in Computer Science from DAVV, Indore, India. His current research interest are Cyber Security, Internet of Things and Machine learning.

Nishchol Mishra is currently an Assistant Professor in School of Information technology Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India. He has completed his M.Tech degree in Computer Science & Engineering from SATI, Vidisha, India . He has received his Ph.D. degree in Computer Science & Engineering from RGPV, Bhopal, India in year 2014. His current research interest are Cyber Security, Data Sciences, Social Media and Analytics.