Social Media Spam Detection

Surya Prakash Mourya
COMPUTER SCIENCE AND ENGINEERING
PDPM IIITDM JABALPUR JABALPUR, INDIA
suryamourya@iiitdmj.ac.in

Raja Nigwal
COMPUTER SCIENCE AND ENGINEERING
PDPM IIITDM JABALPUR JABALPUR, INDIA
rajanigwal@iiitdmj.ac.in

Pradeep Kumar
COMPUTER SCIENCE AND ENGINEERING
PDPM IIITDM JABALPUR JABALPUR, INDIA
pradeepkumar@iiitdmj.ac.in

***Abstract*—As the use of the Internet is increasing, people are connected virtually using social media platforms such as text messages, Facebook, Twitter, etc. This has led to increase in the spread of unsolicited messages known as spam which is used for marketing, collecting personal information, or just to offend the people. Therefore, it is crucial to have a strong spam detection architecture that could prevent these types of messages. Spam detection in noisy platform such as Twitter is still a problem due to short text and high variability in the language used in social media. In this paper, we propose a novel deep learning architecture based on Convolutional Neural Network (FF) and Long Short Term Neural Network (LSTM)**

***Index Terms*—Spam detection Deep learning Sequential Stacked FF-LSTM FF LSTM**

I. Introduction

Online Social Networks (OSNs), like Twitter and Facebook, have become integral to people's daily life in the last few years. Users spend vast time in OSNs making friends with people who they are familiar with or interested in. After the relation is built, users can view messages, usually something interesting or recent activities shared by friends they are connected to, in the terms of tweets, wall posts or status updates. Twitter, which was founded in 2006, has become one of the most popular microblogging service sites. Nowadays, 200 million Twitter users generate over 400 million new tweets per day. Due to the increasing popularity of Twitter, spammers are turning into the fast-growing platform. Twitter spam, which is referred as unsolicited tweets containing malicious links that directs victims to external sites containing mal ware downloads, phishing, drug sales, or scams, etc, has already affected a number of users. In April of 2014, Twitter was flooded with an avalanche of spam tweets that were sent by loads of compromised accounts As a result, the research community, as well as Twitter itself, has proposed a number of spam detection schemes to make Twitter a spam-free platform. For instance, Twitter has applied some rules to suspend accounts if they behave abnormally. Those accounts, which are frequently requesting to be friends with others, sending duplicate contents, mentioning other users or posting URL-only contents, will be suspended by Twitter [4]. At the same time, researchers have proposed innovative mechanisms to detect Twitter spam. The increase in user base of Apps like Twitter has led to the generation and exchange of huge volume of date on the web [5]. While the power to rapidly share information attracts individuals and companies, it also acts as an attraction force for people sending unwanted and unsolicited messages over the network. This type of message is known as a spam message. They are generally marketing, fraud or offensive messages that try to take advantage from the receivers. Spam Detection started with manual filtering of messages, followed by simple filtering rules that could detect a message with some known properties. Automatic spam detection started with the use of traditional machine learning methods that were used to create spam detection model.The latest development in the field of classification is through Deep Learning Technologies. They are known to demonstrate a remarkable performance in the field of Natural Language Processing (NLP) [1]. In this paper, a novel architecture is introduced that combines Convolutional Neural Network (FF) and Long Short Term Memory (LSTM) neural language models and proposed a hybrid deep learning architecture which we have named as Sequential Stacked FF-LSTM model (SSCL) for the spam classification. It offers more diverse text representation which is further enhanced by training the network on the top of pre-trained vectors. In the SSCL model, the text is converted to the vector form with the help of word2vec. This process is also supported by the use of semantic dictionaries like WordNet and ConceptNet. These dictionaries help in extracting the closest semantic word for a given word, for which no corresponding word vector could be found using word2vec. FF perform the task of extracting the most important n-gram features from the text sentence and in the following layer LSTM works upon these features sequences by capturing longterm dependencies. The proposed architecture is evaluated on spam classification tasks and compared with traditional machine learning models as well as individual FF and LSTM architectures. The work presented here is an extension of our previous work. The experimental results show that SSCL architecture produces better results when compared with several benchmark models as well

Identify applicable funding agency here. If none, delete this.

as the FF and LSTM models when used individually. The organisation of the paper is as follows. Section 2 presents the problem statement addresses in this paper. Section 3 presents the related work. Section 4 gives the detail about the proposed SSCL approach that includes two main stream algorithms: FF and LSTM along with the different embedding techniques. Section 5 is the experiments and results section with error analysis. Section 6 concludes the paper with summarizing the contribution.

II. Problem definition

A. Maintaining the Integrity of the Specifications

Spam detection is considered as a NLP classification problem using machine learning algorithms. The problem of spam detection addressed in this paper is described as follows. Given a collection of M annotated text messages x1, x2, x3...xn, for (i = 1 to M) with annotations mi. The annotations mi indicate the i text message is a spam ham message. Spam detection is considered a classification problem, in which for a given short text message, the objective is to classify as spam or ham.

III. RELATED WORK

The severe spam problem in Twitter has already drawn researchers' attention. They have proposed a number of ways to tackle this problem. Some preliminary works used heuristic rules to detect Twitter spam. used a simple algorithm to detect spam in robotpickupline (the hashtage was created by themselves) by using three methods: suspicious URL searching, matching username pattern and keyword detection. simply removed all the tweets which contained more than three hashtags to filter spam in their dataset to eliminate the impact of spam for their research. Other works applied machine learning algorithms for Twitter spam detection. [2], made use of account and content features, such as account age, the number of followers/followings, URL ratio and the length of tweet, to distinguish spammers and non-spammers. These features can be extracted efficiently but also fabricated easily. Thus, some works [7] proposed robust features which rely on the social graph to avoid feature fabrication. Song et al. extracted the distance and connectivity between a tweet sender and its receiver to determine whether the tweet is spam or not . While in [7], Yang et al. proposed more robust features based on the social graph, such as Local Clustering Coefficient, Betweenness Centrality and Bidirectional Links Ratio. Such features were proved to be more discriminative than the features in previous works. However, collecting these features are very time-consuming and resource-consuming, as the Twitter social graph is extremely huge. Consequently, these features are not suitable for online detection.

Initially, spams started spreading with email spam known as unsolicited bulk email (UBE) or unsolicited commercial email (UCE). Further, the SMS being a very cost effective method used for sending individual messages to the prospective clients, has a higher response rate as compared to email spam. Along with emails and SMS, social networking like Twitter [3], Facebook, instant messenger like WhatsApp etc. are also contributing to a major chunk of spam over the network. Spam detection is a tedious task if no automatic filter is installed at the receiving end. One of the initial classifiers is rule based filtering in which the rules are more formally written and can be deployed to a wide area of clients. It includes a set of predefined rules that are applied to an incoming message and the message is marked as spam if the score of the test exceeds the threshold specified. Work in the area of spam and spammer detection on Twitter is already on progressive path. Miller et al. [4] used the content based features like hashtags, mentions etc along with twitter stream clustering methods to identify spammers. Wu et al. [4] used a unified approach consisting of spammer and spams to refine the spam detection results.The same approach was followed by Stringhini et al. [6] in which the social behavior of spammers was combine with the content based features along with SVM to detect the spammers on Twitter.

IV. Deep learning techniques for spam detection

Deep Learning Models have proven their capabilities in the area of NLP tasks like distributed word learning, sentence and document representation [2], parsing [3], statistical machine translation [7], sentiment classification [1, 5], etc. External domain knowledge is important in case of sentence representation through NN in NLP text classification. In many recent researches, the input word sequence and syntactic parse tree are used for creating the text representations while modelling RNN and FF. The pooling feature of FF add power to the model by capturing local as well as high-level features from consecutive context windows. Collobert et al. [5] applied max pooling operation to successive windows to extract global features. Kim [8] also proposed a FF architecture with a number of filters with various window size and two different channels of word vectors. Tao et al. [7] applied tensor-based operations on concatenated word vectors in the convolutional layer. Mou et al. [3] applied convolutional models on sentences having hierarchical structures. Many variants of RNN have been proposed which is able to handle variable length input sequences [4]. Tai et al. [7] changed LSTM structure to model tree-structured topologies by stacking both FF and LSTM in a sequential manner and achieved promising results for semantic sentence modeling. This combination of two deep learning structures can also be seen in some computer vision tasks like image caption [2] and speech recognition [3]. Most of these models use multi-layer FFs or train FFs and RNNs separately or throw the output of a fully connected layer of FF into RNN as inputs.

In this research, our focus is on spam classification in social media using deep learning based approaches. The main concept of the paper is that we add new semantic layer just before the embedding layer that incorporates semantic information. We present three different architectures to achieve this, firstly, we use semantic layer with convolutional neural network, we call this model as Semantic Convolutional Neural Network (SFF). Secondly, we use the semantic layer with the LSTM neural network, this model is named as Semantic long short

term memory (SLSTM). Finally, we present a hybrid of these two approaches, in which we combine the FF and LSTM to make final predictions, we name this hybrid model as Sequential Stacked FF-LSTM Model (SSCL). The details of these models are desbribed in subsequent subsections.

A. Semantic convolutional neural network

In deep learning, a convolutional neural network (FF, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery.

FFs use a variation of multilayer perceptrons designed to require minimal preprocessing.[1] They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics.[2][3]

Convolutional networks were inspired by biological processes[4][5][6][7] in that the connectivity pattern between neurons resembles the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field. SFF is a convolutional neural network with an additional layer into its semantic representation. The initial text along with its word2vec based representation is enhanced using WordNet and ConceptNet knowledge-bases, and further, fed into the traditional convolutional neural network

B. Semantic long short term memory

Long short-term memory (LSTM) is an artificial recurrent neural network, (RNN) architecture[1] used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections that make it a ”general purpose computer” (that is, it can compute anything that a Turing machine can).[2] It can not only process single data points (such as images), but also entire sequences of data (such as speech or video) The LSTM is a variant of recurrent neural network with an enhanced semantic representation of the words. The only difference with the SFF model is that in this approach, we use LSTM neural network in place of FF network to evaluate the robustness of the proposed approach. Further, we also present an approach which combines both of these two approaches SFF and SLSTM that we explain in next subsection.

V. Methodology

A. Data Collection

Data is collected from uci data repository.For the evaluation of the SSCL architecture, two datasets have been used: SMS spam and Twitter dataset. First, the SMS Spam dataset, which is available in UCI repository.3 The dataset consists of 5,574 English, and non-encoded SMS text messages which is a mix of pre-labelled 4,827 ham and 747 spam messages.We have also used tweets from the Twitter dataset which were scraped from public live tweets. These tweets are manually labelled as ham or spam by two annotators ignoring the tweet if there is any conflict among them.

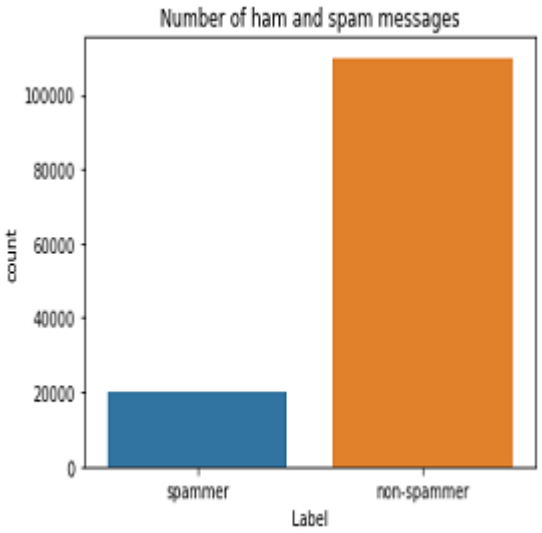

Fig. 1. NO of spam and non-spammers

B. Data pre-processing

Data preprocessing is an important step in the data mining process. The phrase ”garbage in, garbage out” is particularly applicable to data mining and machine learning projects.

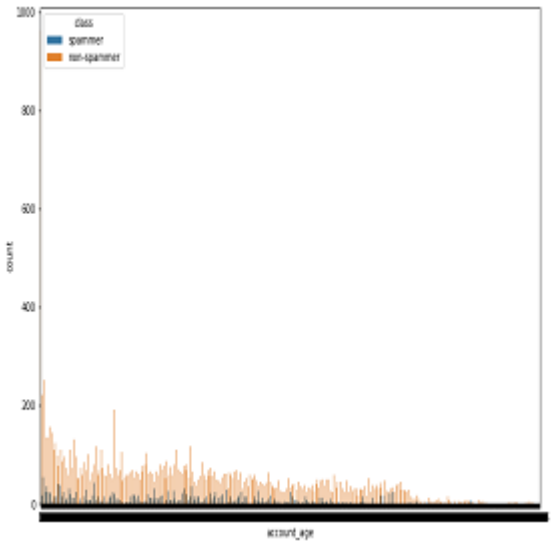

Fig. 2. NO of spam and non-spammers

C. Model training

Random forest, navie byes, decision tree are used to detect spam. later, FF and LSTM are used. model is train and hybrid model is created to get best result.
SVM classifiers
Accuracy for training : 99.32820512820
Accuracy for testing : 99.23846153846154
Decision Tree
Accuracy for testing : 99.63846153846154

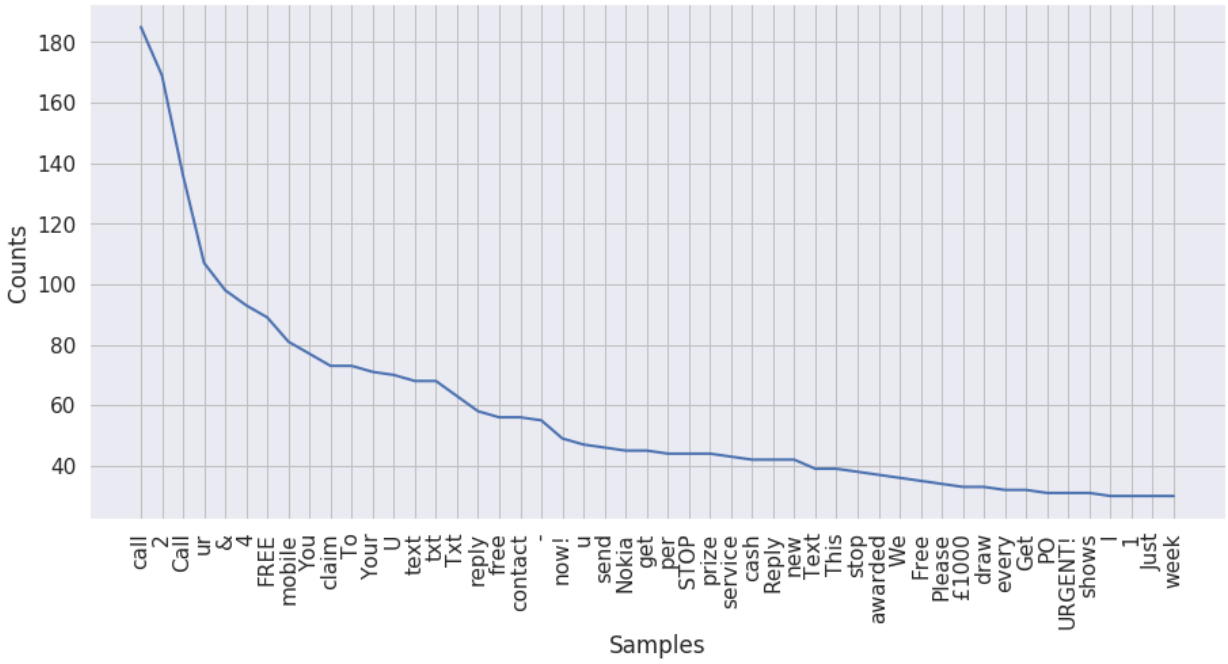

Fig3. plot of 50 most commonly used words in SPAM comment

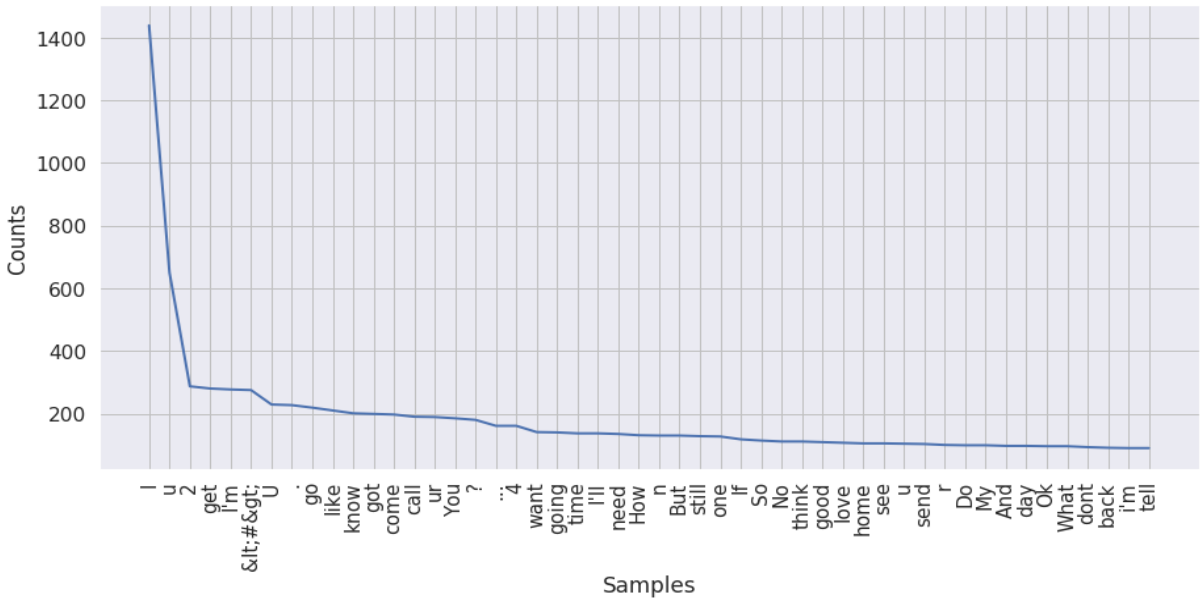

Fig4. plot of 50 most commonly used words in HAM comment

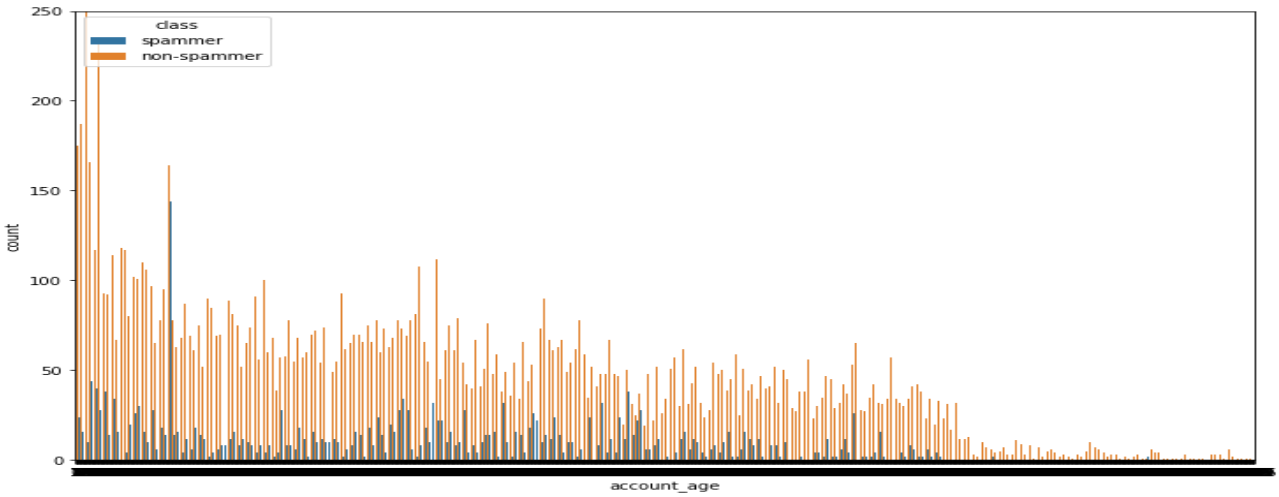

Fig 5. Spam and non Spammer According to account age

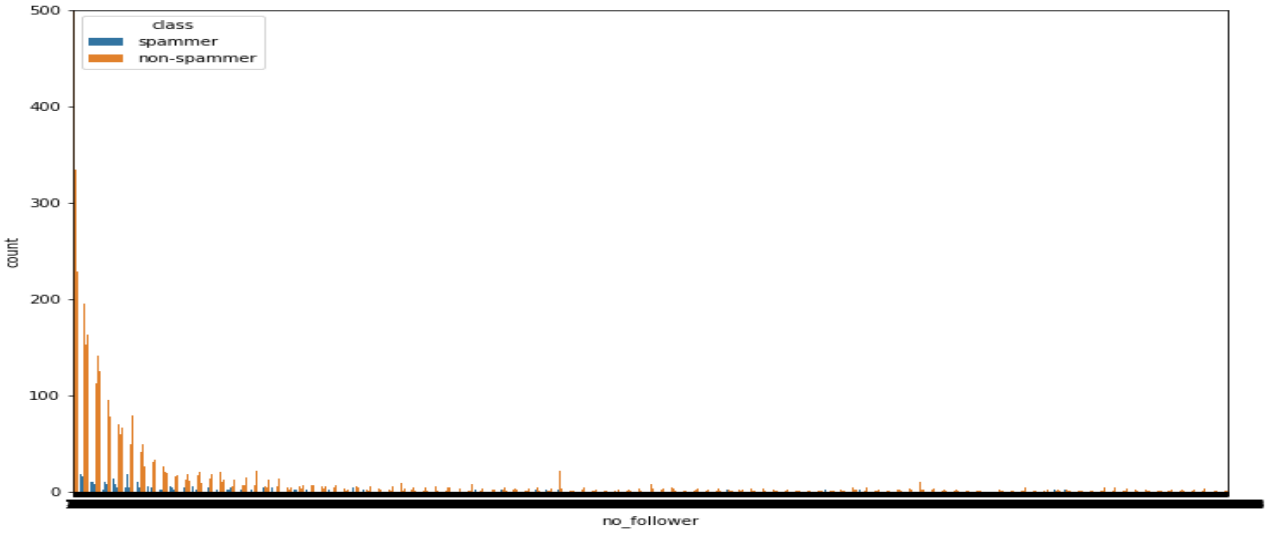

Fig6. Count plot for ' no follower ' feature, based on Spam/ Ham tweet label

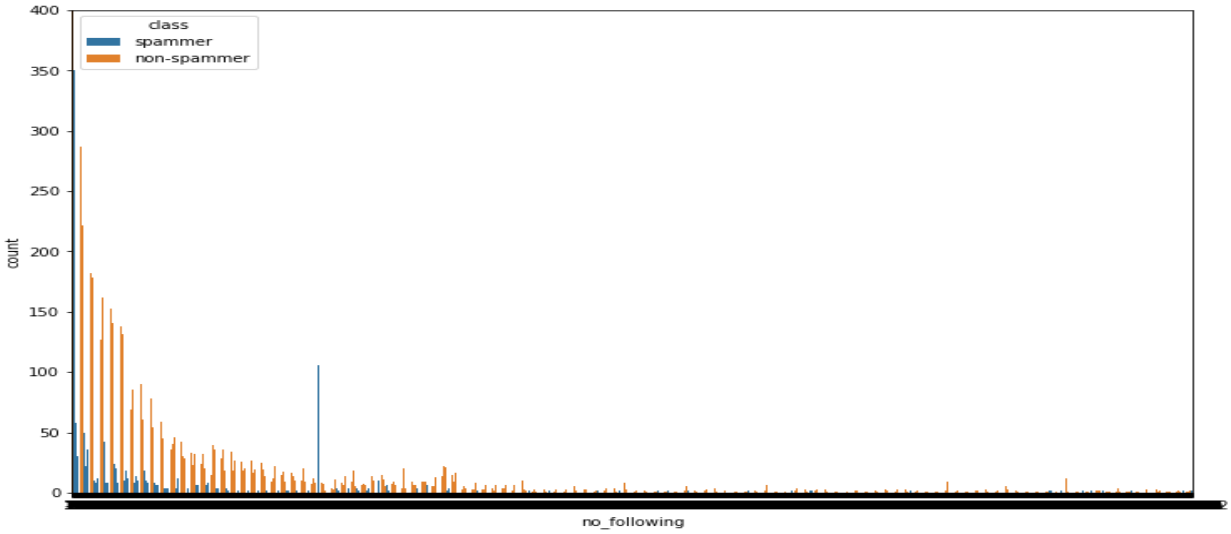

Fig 7. Count plot for ' no_following ' feature, based on Spam/ Ham tweet label

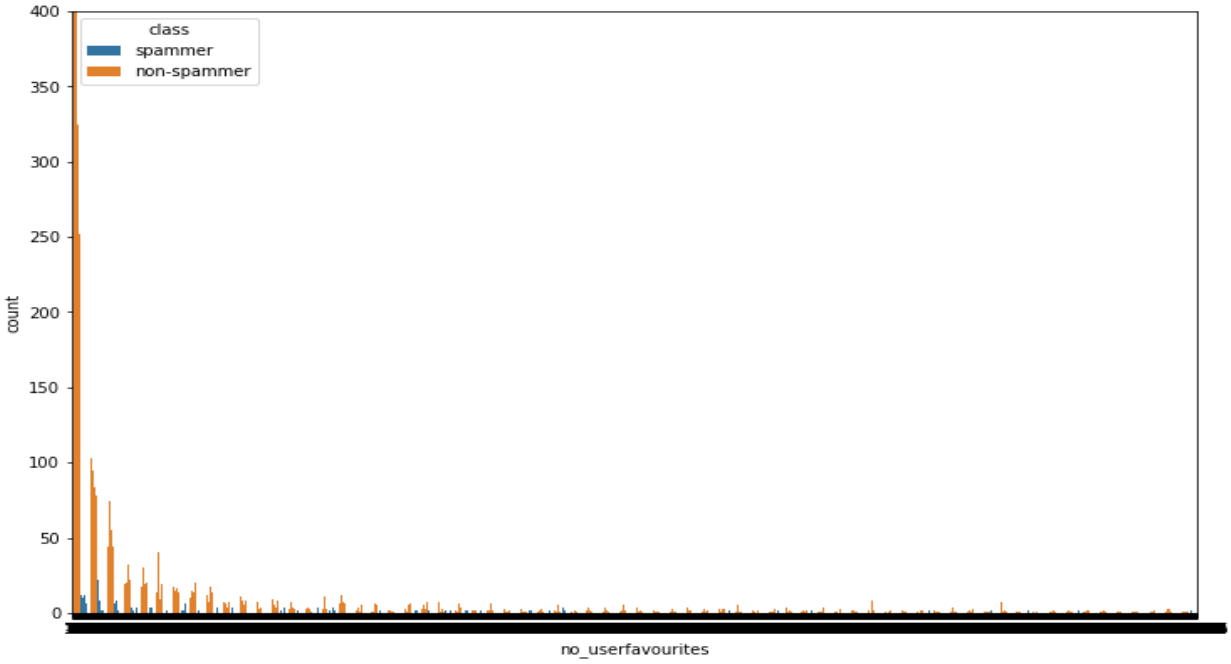

Fig 8. Count plot for ' no_userfavourites ' feature, based on Spam/ Ham tweet label

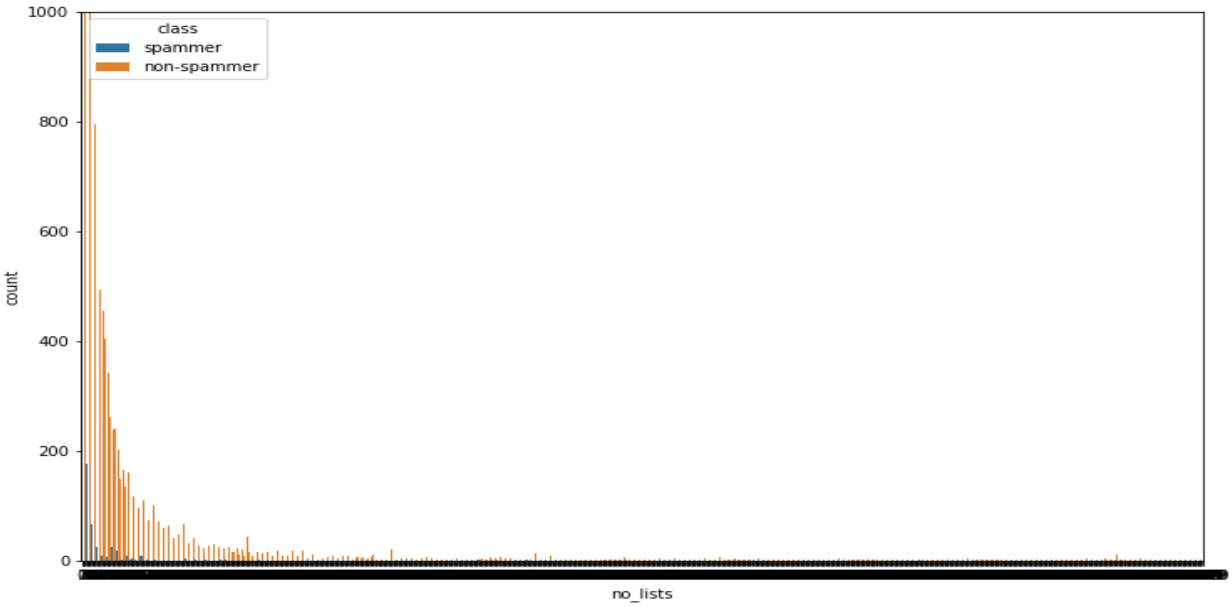

Fig 9.Count plot for ' no_lists ' feature, based on Spam/ Ham tweet label

Naive Bayes
Accuracy for training : 92.14700854700854
Accuracy for testing : 92.92307692307692
Random forest classifier
Accuracy for testing : 99.63076923076923

D. Natural Language And Text Processing in SMS Dataset

1.Importing and parsing text message 2.Filtering stop words 3.Visualising most used words

RESULT

LSTM applied on weibo website dataset and natural language and text processing applied in sms. Spam detection in social media is very challenging problem mainly due to short-text, use of very noisy and irregular language on the social media. In this paper, we focus on spam detection in social media using deep learning approach. Existing successful approaches mainly focused on long email messages but spam detection in short noisy text such as in Twitter is very challenging due to sparseness problem and irregularity in the language used on social media. We proposed a deep learning based approach consisting of FF and LSTM neural architectures. Accuracy: 0.978 which is increase.

ACKNOWLEDGMENT

I respect and thank PDPM IIITDM JABALPUR and Dr. Ayan Seal and Dr. Kusum kumari Bharti for providing me an opportunity to do the project work and giving all support and guidance. I am highly indebted to Dr. Ayan Seal and Dr kusum kumar Bharti for his guidance and constant supervision as well as for providing necessary information regarding the project.

REFERENCES

[1] Agarwal, B., Mittal, N.: Sentiment analysis using conceptnet ontology and context information. In: Prominent Feature Extraction for Sentiment Analysis. Springer. https://doi.org/10.1007/978-3-319- 25343-5 (2016)
[2] H. Tsukayama, "Twitter turns 7: Users send over 400 million tweets per day," Washingon Post, March 2013.
[3] H. Tsukayama, "Twitter turns 7: Users send over 400 million tweets per day," Washingon Post, March 2013.
[4] F. Benevenuto, G. Magno, T. Rodrigues, and Y. Almeida, "Detecting spammer on twitter," in Seventh Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, July 2010.
[5] D. Goodin, "Mystery attack drops avalanche of malicious messages on twitter," Ars technica, Aprial 2014. [Online]. Available: http://arstechnica.com/securityI20l4/04/mystery-attackdrops-avalanche-of-malicious-messages-on-twitterl
[6] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. 12, 24932537 (2011)
[7] K. Thomas, C. Grier, J. Ma, Y. Paxson, and D. Song, "Design and evaluation of a real-time uri spam filtering service," in Proceedings of the 2011 IEEE Symposium on Security and Privacy, ser. SP '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 447-462.
[8] E. Tan, L. Guo, X. Zhang, and Y. Zhao, "Unik: Unsupervised social network spam detection," in Proceedings of 22nd ACM International Conference on Information and Knowledge Management, San Fransisco, USA, October 2013.