www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
Vol. 01, Issue 02, Nov. 2021, pp: 5-9

e-ISSN : xxxx-xxxx
Impact Factor : xxxx

SMS SPAM MESSAGE DETECTION

Devendra Bist*1, Rajan Vyas*2, Chetesh Tiwari*3

*1,2,3Dept. of IT, Dr. Akhilesh Das Gupta Institute Of Technology & Management, India.

ABSTRACT

The popularity of SMS has also given rise to SMS Spam, which refers to any irrelevant text messages delivered using mobile networks. They are severely annoying to users. spam messages can lead to loss of private data as well. Spam SMSes are unsolicited messages to users, which are disturbing and sometimes harmful.In this paper, We used a public SMS Spam dataset, which is not purely clean dataset. The data consists of two different columns (features), such as context, and class. The column context is referring to SMS. The column class may take a value that can be either spam or ham corresponding to related SMS context.Before applying any supervised learning methods,we applied a bunch of data cleansing operations to get rid of messy and dirty data since it has broken and messy context. To fix the problem of data leakage we applied pipeline class to each classifier and used hyperparameter tuning from GridSearchCV to increase the efficiency of the model.

I. INTRODUCTION

In this day of age, everyone owns a smartphone and the availability of computers and portable laptops is increasing and Since the cost reduction of Short messaging services(SMS), these spam messages have increased exponentially over the years. It was studied that spam was most commonly present in emails than compared to mobile SMS as sending spam SMS to mobile is a lot more costly compared to sending one in email, But mobile phones being easy to use have made phishers consider SMS messages as a better method, phishers can buy multiple devices for more profit then phishers send malicious URL through SMS which redirects the user to address which inturns steal their sensitive information. Spam messages can come from any part of the globe with China being number 1 in sending the most spam and Vietnam as a close second. The continuous spreading of this kind of problem and less effective security measures has inspired many researchers around the world in the development of a set of techniques to help prevent it efficiently. Many users are still unaware of protection mechanisms, thereby, making their mobiles prone to cyber-attacks. India has set up an NCPR registry, which has to some extent reduced spam calls but does not filter spam SMS. Therefore we have come up with a better classification of SMS spam in our study to tackle this problem.

II. LITERATURE REVIEW

Nilam NurAmir Sjarif et al . For instance, build a spam detection model from the UCI machine learning repository and applied TF-IDF over several supervised learning algorithms.

Mehul Gupta et al Did a comparative study of spam SMS detection using machine learning classifiers and using Cumulative Accuracy Profile (CAP) Curve which is a much more robust and better method to compare and assist machine learning classifiers.

O. O. Abayomi-Alli et al did a critical analysis on existing SMS spam machine learning detection models they researched on content, non-content, collaborative, and adaptive based filters and summarising the problems with these filters.

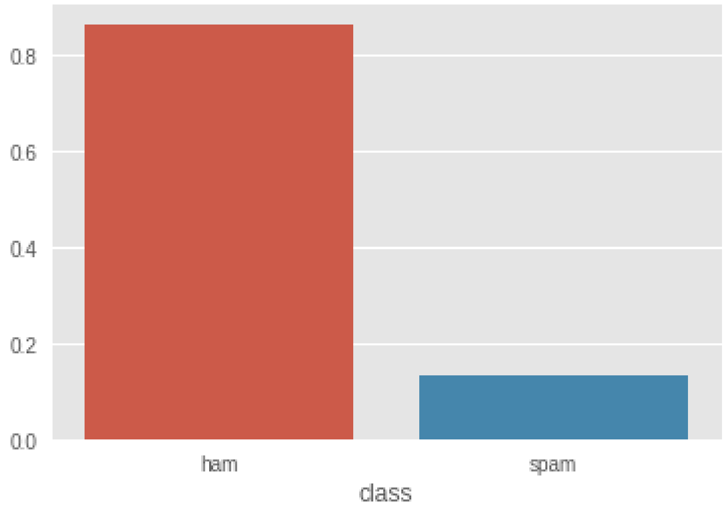
Saeid Sheikhi et al Made An Effective Model for SMS Spam Detection Using Content-based Features and Averaged Neural Network and getting the best results with their neural network model up to 0.988% accuracy and 0.9929% F-measure rate respectively.

Sakshi Agarwal et al did spam detection on Indian messages they analyzed different machine learning classifiers on a large corpus of SMS messages for Indian people. They applied various features on different classifiers and plotted the results with the Support vector machine having the highest accuracy out of all the other classifiers used.

III. METHODOLOGY

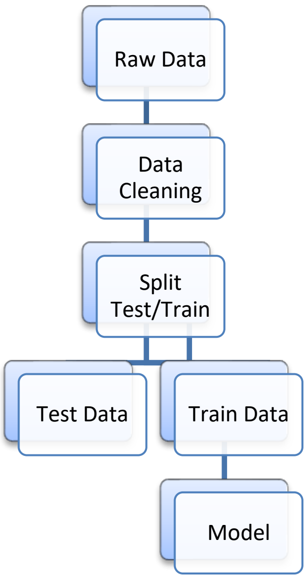
- Dataset Description:

A collection of 425 SMS spam messages was extracted from the Grumbletext Web site. A subset of 3,375randomly chosen Mobile SMS of the NUS SMS Corpus (NSC), having 10,000 dataset legitimate messages collected for research at the Department of computing at the National University of Singapore. The messages mostly originate from Singaporeans and from students who attending the University. This dataset is not purely clean. This dataset consist of two different columns one column contains the message and another column contains the class that the message is spam or ham.



- Data Prepossessing:

Different preprocessing approaches have been used in this project. Following is the list of them:



PART 1 classifying SMS using supervised learning methods

- In this part, we used the pipeline approaching to apply a bunch of different operations respectively on the data. The class of SMSClassfication has 3 different pipelines corresponding to different ML algorithms, such as;
- Naive Bayes (NB),
- SVM,
- Random Forest Tree (RFT).
- Using spacy:

We create tokens and lemmas of the dataset using Spacy library. Spacy is an open-source library that is developed in python and cython for use in advanced natural language processing. The spacy library provides the best way to do something than the NLTK.

- Bag of Words:

As the dataset is messy and Machine learning algorithms prefer clean and well-defined input and outputs. The machine learning algorithm cannot work with raw data directly, the text should be converted into numbers. Specifically, Vectors of numbers. Bag of words is the representation of an occurrence of a word in the documents.

- Term Frequency Inverse Document Frequency:

TF-IDF (Term Frequency Inverse Document Frequency) is the technique to compute the weight of each word which signifies the importance of the word in the document.

$$\mathbf{tf}(t, d) = \frac{f_d(t)}{\max\limits_{w \in d} f_d(w)}$$

$$\mathbf{idf}(t, D) = \ln\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

$$\mathbf{tfidf}(t, d, D) = \mathbf{tf}(t, d) \cdot \mathbf{idf}(t, D)$$

$$\mathbf{tfidf}'(t, d, D) = \frac{\mathbf{idf}(t, D)}{|D|} + \mathbf{tfidf}(t, d, D)$$

$$f_d(t) := \text{frequency of term t in document d}$$

$$D := \text{corpus of documents}$$

- Singular Value Decomposition:

SVD (Singular Value Decomposition) is the technique to reduce the dimension of the data. It is one of the most useful algorithm for dimension reduction.

A = S∑U^T

where, S → Eigen Vectors of A^TA

∑ → Diagonal Matrix of singular values A^TA

U^T→ Eigen Vectors of AA^T

- Pipelining :

A machine learning pipeline is a way to automating the ML workflow by enabling data to be translated and correlated into the model and achieve required outputs.

- Splitting Dataset:

The dataset was split into two parts – One is train dataset which used to train the dataset and the test dataset which is used to find the accuracy of the model.

- Classifiers:

Followings are the classifier which is used in this experiment:

1. Naïve Bayes:

Naïve Bayes is the classification algorithm based on the Bayes theorem. Bayes theorem finds the probability of the occurring an event when the probability of another event is already given.

2. Random Forest:

Random Forest Classifier based on an ensemble of a large number of the individual decision tree. Each decision tree in random forest gives its prediction and the result got from the majority of classes will be the final output of that parameters.

3. SVM (Support Vector Machine):

SVM is the supervised learning model, and one of the most robust prediction method. In SVM
Two classes are separated by a boundary known as a hyperplane. The SVM algorithm aims to find the best decision boundary (Hyperplane) that can segregate n-dimensional space into classes so that we can easily put the new data point in the classes.

Part 2: Classifying SMS by using Deep Learning with RNN (LSTM)

- In this part, we applied deep learning.
- During this experiment,
  - To build an ML model based on Deep Learning, we used Keras API and its backend is Tensorflow.
  - To apply the NLP technique, we used Spacy libs rather than NLTK, again.
  - To create word2vec model and embedding vector before applying deep learning, we used Gensim libs instead of using TF-IDF. We could use Google's, GloVe's, Spacy's pre-trained vectors. However, we built our word2vec model since we have domain-specific data based on SMS.
- We built a deep learning network by using the layers, respectively. We connected each other layer as you see in the graph, below.
  - Embedding Layer
  - Dense Layer
  - LSTM for RNN Layer
  - Dense Layer
- After we split the message into tokens by using Keras' tokenizer, we plotted the CDF graph for the frequency of unique words.
- According to that graph,
  - the unique words appear less than 50 times in 95% of the corpus.
  - the unique words that appear less than 1 time in 50% of the corpus.
  - the unique words that appear less than 4 times in 75% of the corpus.

- Hyperparameter Tuning:

Hyperparameter Tuning is a technique for choosing a set of optimal parameters for a model from the different sets of the parameter. In this project, we used the GrideSearchCV to find the best parameters for our model. So that we can accurately predict the classes.

- Measuring Matrix:

We used the confusion matrix to find the best classifier for this project. It describes the performance of a model in tabular form.

	PREDICTED: NO	PREDICTED: YES
ACTUAL: NO	TN	FP
ACTUAL: YES	FN	TP

The followings are the terms described in the confusion matrix:
True Negative: When it's actually no, how often does the model predict no?
False Negative: When it's actually Yes, how often does the model predict no?
False Positive: When it's actually no, how often does the model predict yes?
True Positive: When it's actually yes, how often does the model predict yes?

IV. RESULT AND DISCUSSION

This section includes the outcomes of the result and describes the performance of each of the classifiers by plotting the table and getting their accuracy, precision, recall, etc. Moreover achieved results were compared to other techniques.

Precision and recall both need to be high to cover two different cases. According to our expectations, the deep learning approach is giving better results in terms of precision and recall metrics.

It has the highest accuracy than the other classification algorithms we applied before since word2vec cares about semantics more and TF-IDF fails to cover that requirement well.

It is notable that out of all the classifiers which used TF-IDF, random forest + TF-IDF had the highest accuracy precision as well as recall while the other two algorithms had lower accuracy as well as precision and recall.

It could be due to the SVM not handling an imbalanced dataset nonetheless SVM+TF-IDF had 97% accuracy.

Algorithm	Accuracy	Precision for spam	Precision for ham	Recall for spam	Recall for ham
TF-IDF+ Naïve Bayes	96.164	1.00	0.96	0.76	1.00
TF-IDF+ Support Vector Machine (SVM)	97.106	0.91	0.98	0.87	0.99
TF-IDF + Random Forest	99.125	1.00	0.99	0.94	1.00
Word2vec + LSTM	99.349	0.98	1.00	0.97	1.00

V. CONCLUSION

The main aim of the project is to develop a model based on the SMS classification in which we identify the SMS into two classes that is ham or spam. We used different text processing techniques like bag of words, TF-IDF, SVD, etc. The result of our paper demonstrates that this approach works pretty well and shows the performance of the different models and the comparison between different models.

VI. REFERENCES

[1] Guillermo Cajigas Bringas, Jose Maria Gomez Hidalgo, Enrique Puertas Sanz, Francisco Carrero García.“Content-based SMS spam filtering”(January 2006)

[2] Gupta, B.B., Tewari, A., Jain, A.K. and Agrawal, D.P., "Fighting against phishing attacks: State of the art and future challenges", Neural Computing and Applications, Vol. 28, No. 12, (2017), 3629-3654

[3] Gudkova, D., M. Vergelis, T. Shcherbakova, and N. Demidova. (2017) “Spam and Phishing in Q3 2017.” https://securelist.com/spam-and-phishing-in-q3-2017/82901/. [Accessed: 10th April 2018].

[4] Mehul Gupta, Aditya Bakliwal, Shubhangi Agarwal, and Pulkit Mehndiratta.(2018) “A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers” Eleventh International Conference on Contemporary Computing 2018

[5] Nilam Nur Amir Sjarif*, Nurulhuda Firdaus Mohd Azmi, Suriayati Chuprat, Haslina Md Sarkan, Yazriwati Yahya, Suriani Mohd Sam. (2019). “SMS Spam Message Detection using Term Frequency-Inverse Document Frequency and Random Forest Algorithm”. The Fifth Information Systems International Conference 2019.