PROTECT PERSONAL PRIVACY AND WASTING TIME USING NLP: A COMPARATIVE APPROACH USING AI

Vivekananda Journal of Research
October, 2021, Vol. 10, Special Issue, Pg No.42-52
ISSN 2319-8702(Print)
ISSN 2456-7574(Online)
Peer Reviewed Refereed Journal

http://www.vips.edu/vjr.php

Arun Velu, [1] Dr. Pawan Whig [2]
Director Equifax and Researcher, Atlanta , USA [1]
Senior IEEE Member, Dean Research, Vivekananda Institute of Professional Studies, New Delhi, India [2]

ABSTRACT

In one of the research Study Joe et. al proposed that 33% People classify SMS Spam as annoying while about 25% wasting time, and nearly 22 % violating personal privacy . Spam is the unwanted irritating email systems to indiscriminately send out spontaneous posts in wholesale. While e-mail spam is most well known, the term is used in other media for similar abuses. Text spam, generally unwanted bulk message with some business interest, is similar in the context of email spams. SMS spam is used to spread phishing links and commercial advertising. Commercial spammers use malware that is illegal in most countries to send SMS spam. Transfer unwanted spam on an endangered machine reduces the danger to the spammer, which complicates the origin of junk mail. The presence of known words, phrases, abbreviations and idioms can greatly influence the detection of SMS spam. The objective of this paper is to analyse and evaluate different classification techniques based on their accuracy, recall and CAP curves. There was a comparison between traditional methods of machine learning and deep learning.

Keywords: machine learning , spammer, malware, SMS

1. INTRODUCTION

Whenever SMS spam reaches the inbox of a user, the mobile phone alerts the user. If the user realizes that the message is unwanted, they are disappointed, and also some mobile telephone storage is carried out by SMS spam. The detection of SMS spam is an important task for identifying and filtering spam text messages. More SMS messages are sent daily so that a operator can correlate newer SMS posts in the situation of previously conventional SMS messages. It is also very challenging. We thus attempt to advance a web-based SMS text junk mail or ham gauge by using the knowledge of artificial intelligence by combining machine learning and data mining.

Electronic mail remains a very economical communication method, but hackers use it as a method for spreading the virus, phishing and malicious code, etc. Email is very useful and convenient to use, but many people misuse it too. Though there has been a steady increase in email use as well as the number of spams, other communication media like Facebook, Twitter, and chat programs. There has been an exponential increase in e-mail service. The average increase in the number of messages sent daily during 2019 was 538.1 million since 2010. Spam volume during 2018 was 89-92%.

Spam emails become one of the most serious issues for email users and affect email users negatively. HAM is a real email to receive and SPAM is a spurious mail to thousands of users with

bad intentions from unreliable sources in bulk. Spam is deliberately sent to users where it is not meant to be received by the recipient. Spam operators attempt by luring them with attractive offers to extract sensitive user data.

Because of the high number of methods for separating HAM and SPAM, a single method which can lead to a very small positive rate is difficult to narrow. Although many methods and algorithms are available to combat Spam Problem, there is no perfect single method or algorithm. During the isolation process, many HAM mails are marked as spam. This situation is called False by misrepresenting HAM as SPAM.

This research paper is organized in various sections. The first area focuses on associated work in the machine and non-machine learning processes used to classify e-mails as HAM or SPAM. The second section provides an insight into the collaborative approach and related work in this field. Spam control. The third section discusses the techniques of various machine learning algorithms to find out comparative analysis and to conclude which algorithm is best in terms of accuracy and other factors.

2. VARIOUS METHODS FOR SPAM DETECTION

ML and Non ML methods

Machine learning and non-machine learning are two major categories for spam filtering. Non-machine learning methods use a spam filter procedure such as White-Listing, Black-List and keyword search . Non-machine teaching is easy to implement and to experiment, so spammers are highly likely to bypass non-machine teaching. However, a strong search for keywords and continuous updating of the whitelist and blacklist may still be more successful.

The approach to machine learning coincides greatly with the characterization of text and thus iterates researchers' interest. Researchers have used many approaches in machine learning such as vector support, memory based learning, Ripper rules-based learning, boosting decision-making processes, hard sets, neural networks and Bayesian classifiers. Most approaches to email classification are based on a single text classification algorithm. Rough theory or rough theory-based methods for email classification are most popular among researchers. Emails are not written as they should be written when the malicious task is used. In such circumstances, emails are difficult to categorise, but the raw theory is quite inaccurate, inconsistent and unprecedented. Some researchers have used data mining methods, such as SVM, neural network, naive Bayesian and theoretical rough-set. The problem, however, was that these methods were employed with a set of rules. Researchers such as Chouchoulas, Zhao and Zhang, Zhao and Zhu have only experimented with spam rule generation keyword frequency.

Cooperation approach

Many solutions are available on standalone servers to classify spam. The solutions on stand-alone servers include Counter Attack, Opt-Outlist and Spam filter. Standalone servers have limited computing power and speed, so a new type of spam mail cannot be filtered by standalone servers. Any new mail type requires the creation and use of new rules. The collaborative approach has the necessary computing power and speed to produce and use new anti-spam rules as required.

Spam prevention

SPAM may be handled in two different ways, i.e. at the place of origin, check and block spam

or during mail receipt. Spammers are aimed at servers that allow another server to use them as an intermediary channel. These low-security unattended servers are called . Because spammers use Botnets and the location of their origin constantly changes, spam can be easily checked at its origin. Some servers are black listing for SPAM spreading or used as a SPAM spreading channel. Since the source is not confident, the mail is called spam directly. On the internet, there are many open proxy servers. Spammers will also spread SPAM on these servers. When an open proxy server is used by spammers, the actual e-mail source is hard to identify. Once the email is sent to the mail server, the second method of verification and classification is HAM and SPAM.

PEAS Representation

PM : Performance Metrics means indicates How systematizes the AI distinguish it's doing?

E: **E**nvironment tells what situation in which the agent relates?

A: Actuators describes the effect prepares AI have on surroundings?

S: Sensors is used to gather AI data starting its surroundings?

The following is a PEAS summary of the classifier's task environment:

Table 1 PEAS representation of AI Problem

Agent	Performance Measure(P)	Environment(E)	Actuators(A)	Sensor(s)
Spam or ham classifier	Correct classification	• Text • Numbers • Alphanumeric-characters	Screen Display(Form)	Keyboard

Description of algorithms

ML processes can be used to recognize junk mail or ham by appropriate scheme to the relevant labels and to make predictions or classifications by means of a trained model as the objective of the project will be to categorize the texts into junk mail or ham, which is the problems of Dual Grouping.

Sack-up or Bagging

One of the important algorithm Bagging practices a modest method that appears repeatedly in statistical analyses — to improve one's estimates by combining many estimates. Bagging constructs n classification trees by sampling bootstrap data and combines predictions in order to yield a concluding prediction.

Random Forest (RF)

RF or forests of haphazard decision are a group technique of learning to classify the regression and to do other tasks by creating a multitude of decision-making trees and delivering the classes that

fashion the individual trees for classes (classification) or mean (regression). Random decision forests have the habit of overfitting their training set for decision-making bodies

Naïve Bavarian (NB)

As their name implies, this algorithm assumes that each of the data sets' variables is "Naïve," i.e. not interconnected. The calculation of each hypothesis is simplified in order to make the calculation tractable. Naïve Bayes or Idiot Bayes are called. Naïve Bayes is an algorithm for viral classification which is used mainly to achieve the dataset's basis accuracy. While the algorithm is naive, it is quicker than substantial data sets for SMS information to be classified on the basis of probability. This reduces the burden on the AI arrangement and retains the productivity efficient and controllable. Because of the lack of a well-trained Bayes classifier, they are relaxed and fast to implement. This allows real data to be tested without outlay much time and change emerging the model. They are prepared to predict when applied.

Classification of Extra Trees

The Random Forest is like this. It produces several trees and splits nodes by using random subsets of characteristics, but it does not bootstrap observations with two key differences (that is, samples with no substitution), and nodes are split in random splits and not in best divisions.

Support Vector Machine (SVM)

Support-vector machines are supervised learning models that analyse data used for classification and regression analysis, with their associated learning algorithms. SVMs are also supported vector networks. In light of a number of training examples, each of which is marked as one category of two, and an SVM training algorithm creates a model that assigns new examples to a single category or another and makes it not likely to be a binary linear classification. SVM is a great way to classify binary data, as SVM categorizes facts by result the best plane and distinguishes facts points from the other class.

KNN

Classification based on KNN K-Nearest in neighbouring countries is a kind of lethargic learning, because it fixes not attempt to shape a general internal model, but only stores training data instances. Classification of the k shall be calculated by simple majority vote of the closest neighbours. This algorithm is easy to implement, robust, bright and useful when there are extensive training data. The drawback is, however, that the value of K must be established and the calculation costs are high, since the detachment of each occurrence from all the training examples must be calculated.

Decision Tree

The division criteria standard lies behind the intellect of any classification of the tab. Decision trees are similarly obtainable to a tree-structured current chart where the instances are classified according to their properties. An instance, the results of the trial labeled by the outlet and the leaf bump represent the class label, represents a node in a Decision Tree.

3. MATHEMATICAL ANALYSIS

Information about dataset used can be given by using following command

data.info

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   v1          5572 non-null   object
 1   v2          5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
```

Bar graph representation of above data is obtained by using following commands

```
#Palette
cols= ["#E1F16B", "#E598D8"]
#first of all let us evaluate the target and find out if our data is imbalanced or not
plt.figure(figsize=(12,8))
fg = sns.countplot(x= data["target"], palette= cols)
fg.set_title("Count Plot of Classes", color="#58508d")
fg.set_xlabel("Classes", color="#58508d")
fg.set_ylabel("Number of Data points", color="#58508d")
```

after executing above commands we get

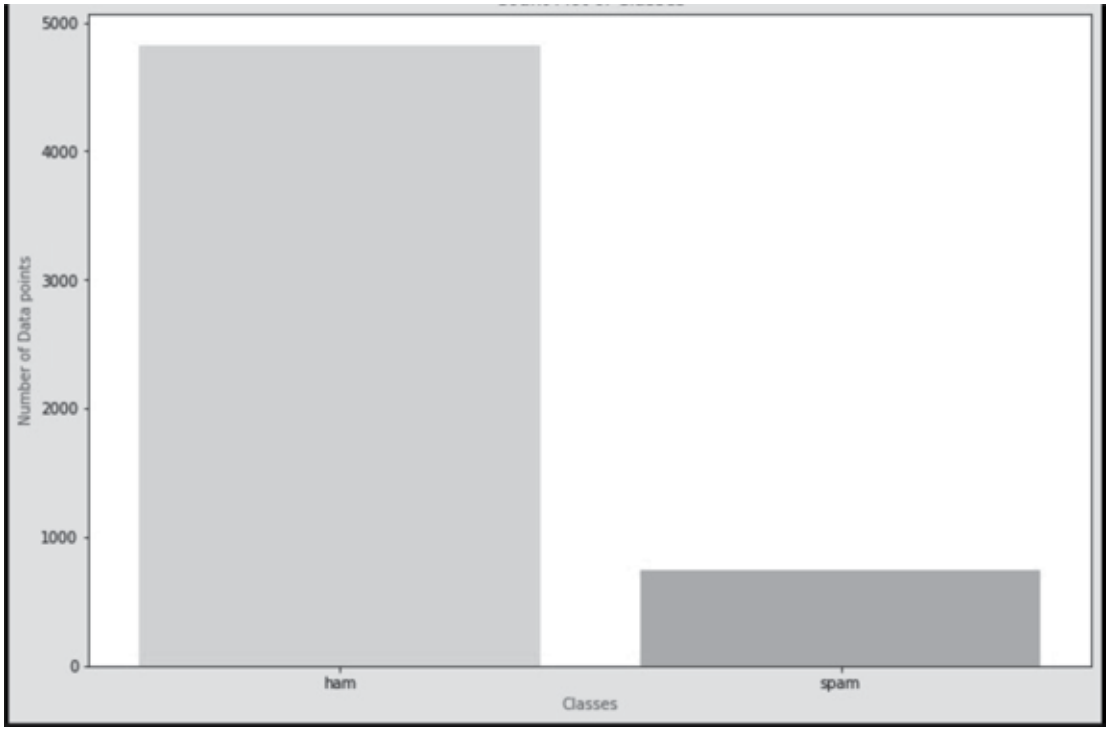

Figure 1 classiffication of data

The description of data is represented by

data[“No_of_Characters”] = data[“text”].apply(len)

data[“No_of_Words”]=data.apply(lambda row: nltk.word_tokenize(row[“text”]), axis=1). apply(len)

data[“No_of_sentence”]=data.apply(lambda row: nltk.sent_tokenize(row[“text”]), axis=1). apply(len)

data.describe().T

	count	mean	std	min	25%	50%	75%	max
No_of_Characters	5572.0	80.118808	59.690841	2.0	36.0	61.0	121.0	910.0
No_of_Words	5572.0	18.501256	13.637056	1.0	9.0	15.0	27.0	219.0
No_of_sentence	5572.0	1.991565	1.501427	1.0	1.0	1.5	2.0	38.0

Top 10 Ham and Spam Words used are given by

#for counting frequently occurence of spam and ham.

count1 = Counter(“ “.join(data[data[‘target’]==’ham’][“text”]).split()).most_common(10)

data1 = pd.DataFrame.from_dict(count1)

data1 = data1.rename(columns={0: “words of ham”, 1 : “count”})

```
count2 = Counter(“ “.join(data[data[‘target’]==’spam’][“text”]).split()).most_common(10)
data2 = pd.DataFrame.from_dict(count2)
data2 = data2.rename(columns={0: “words of spam”, 1 : “count_”})
```

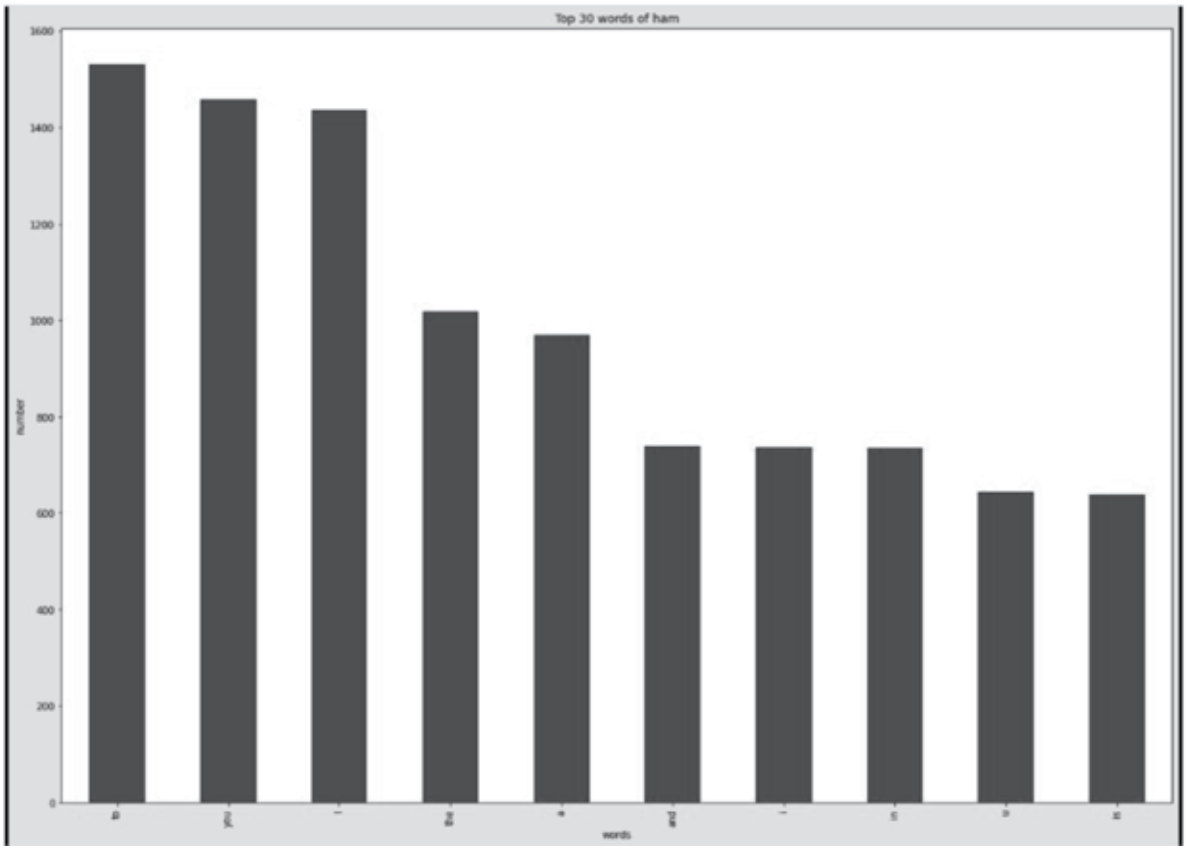

Figure 2 Frequency of ham words

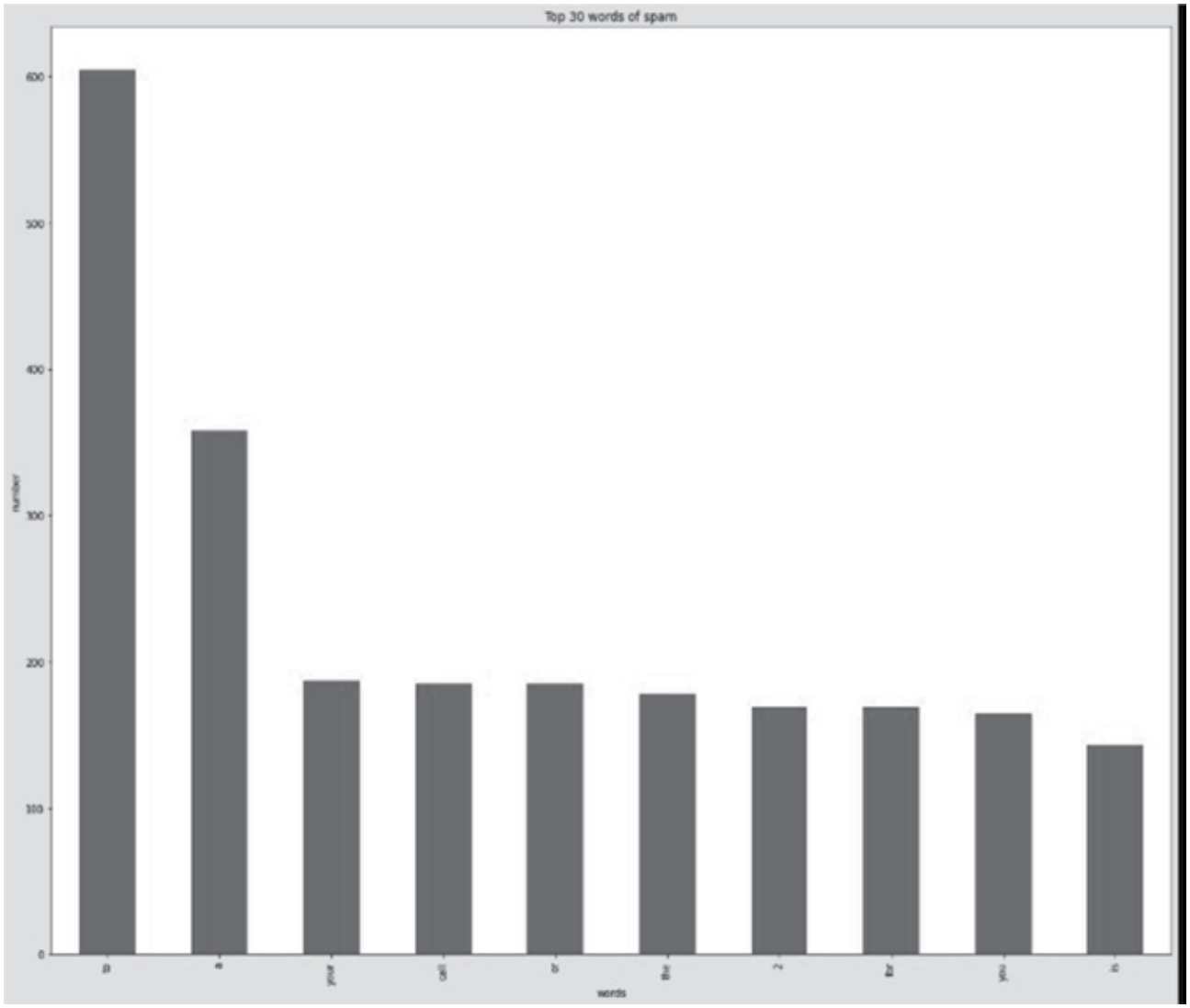

Figure 3 frequency of spam words

4. ANALYSIS AND ASSESSING

The following metrics were used to determine certain evaluation metrics:

- True Positive (TP) - the number of test cases correctly classified;
- True Negative (TN) - the number of test cases correctly rejected from the main class.
- False Positive (FP) - the number of test cases that are incorrectly rejected from the main class;
- False Negative (FN) - the number of test cases that are incorrectly classified to the main class.

We used accuracy as the primary evaluation criteria for our classifiers because it is an intuitive metric with a simple interpretation: simply counting correctly classified messages.

Important Matrices

To get the required information to validate any proposal or comparison, it is significant to select the precise performance metrics for setup. We consider the following known metrics in order to analyse

and compare the detection capacities of the classifiers considered:

Accuracy

Accuracy is the most intuitive performance measurement and simply a proportion with all the observations that have been accurately predicted. One could believe that if it is highly accurate, our model is best. Yes, accuracy is only a big step if the values of false positive and negative are almost identical, but only when you have symmetrical data sets. Therefore, you need to look at other parameters to assess the performance of your model.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

The ratio of correctly predicted positive observations to the total predicted positive observations is known as precision. The question that this measure answers is how many of the passengers who were labelled as having survived really did. The low false positive rate is related to high precision.

Precision =

Recall

It is also called sensitivity and defined as the correctly predicted ratio of positive observations to all actual class observations.

$$\text{Recall} = \frac{TP}{TP+F}$$

Table 2 Comparisons among various algorithm

Algorithm Used	True Positive	False Negative	False Positive	True Negative
Naive Bayes	1425	15	14	218
SVM	1303	141	136	92
KNeighbors	1439	0	233	0
Decision Tree	1421	27	18	206
Extra Tree	1434	25	5	208
Random Forest	1437	32	2	201
AdaBoost	1417	36	22	197
Bagging	1422	27	17	206

Table 3 Comparison in terms of Accuracy ,Precision and Recall

Algorithm	Accuracy	Precision	Recall	AR	Positive
Naïve bayes	98.7	0.98	0.93	0.92	0.94
SVM	98.9	0.96	0.98	0.93	0.98
KNeighbors	98.7	0.94	0.92	0.96	0.98
Decision Tree	98.5	0.97	0.96	0.96	0.94
Random Forest	98.0	0.95	0.98	0.93	0.95
AdaBoost	98.1	0.94	0.99	0.98	0.97
CNN	99.2	0.99	0.94	0.94	0.89
Bagging	98.2	0.94	0.99	0.99	0.99

5. RESULT AND CONCLUSION

This research studies focused on debating about how to protect personal privacy and wasting time with the application of AI enabled technology and testing ML techniques for junk mail or Junk SMS detection. Several algorithms compared with different classifiers to see which one was the best. The results of our classifier evaluation show that the CNN Classifier achieves the highest accuracy of 99 percent and about 98 percent, respectively, for the given dataset, with AR values of 0.99 . While CNN has been widely used in image classification, it has shown substantial improvements over predictable classifiers and achieves the maximum accurateness among them for word-based data as well. This achievement by CNN greatly expanded the research area of its presentation to text-related classification issues, such as review classification and sentiment prediction. SVM and NB, as predicted, perform well among conventional classifiers, coming close to CNN for both datasets. Noteworthy outcomes have been got as a result of this work, indicating that this study can be applied in the real world to detect spams SMS. This research Study can be very useful for the researchers working in the same field.

Future Scope

This research study can be helpful for the researchers working in the same field. Some New Machine learning or deep learning algorithm can be applied in order to get more accurate result.

REFERENCE

[1] H. Najadat, N. Abdulla, R. Abooraig, and S. Nawasrah, “Mobile SMS Spam Filtering based on Mixing Classifiers,” International Journal of Advanced Computing Research, vol. 1, 2014

[2] J. Deng, H. Xia, Y. Fu, J. Zhou, and Q. Xia, “Intelligent spam filtering for massive short message stream,” COMPEL-The international journal for computation and mathematics in electrical and electronic engineering, vol. 32, pp. 586-596, 2013.

[3] K. Yadav, P. Kumaraguru, A. Goyal, A. Gupta, and V. Naik, “SMSAssassin: Crowdsourcing Driven Mobile-based System for SMS Spam Filtering”, HotMobile’11 Proceedings of the 12th Workshop on Mobile Computing Systems and Applications, pp. 1-6, 2011.

[4] B. S.-E. Kim, J.-T. Jo, and S.-H. Choi, “SMS Spam Filtering Using Keyword Frequency Ratio,” International Journal of Security and Its Applications, vol. 9, pp. 329-336, 2015.

[5] T. M. Mahmoud and A. M. Mahfouz, "SMS Spam Filtering Technique Based on Artificial Immune," IJCSI International Journal of Computer Science Issues, vol. 9, 2012.

[6] Y. Yang and S. Elfayoumy, "Anti-Spam Filtering Using Neural Networks and Bayesian Classifiers", Proceedings of the 2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation, Jacksonville, FL, USA, June 20-23, 2007.

[7] David Ndumiyana, Munyaradzi Magomelo, and Lucy Sakala, "Spam Detection using a Neural Network Classifier," Online Journal of Physical and Environmental Science Research, vol. 2, issue 2, pp. 28-37, April 2013.

[8] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the Study of SMS Spam Filtering: New Collection and Results (preprint)," Proceedings of the 11th ACM symposium on Document engineering, Mountain View, California, USA, pp. 259-262, 2011.

[9] Pawan Whig and S. N Ahmad," Modelling and Simulation of economical Water Quality Monitoring Device ", Journal of aquaculture & Marine Biology, 2016, Vol.4,Issue 6, pp.1-6**(Scopus)**. **ISSN: 2378-3184**

[10] Pawan Whig and S. N Ahmad," Controlling the Output Error for Photo Catalytic Sensor (PCS) Using Fuzzy Logic ", Journal of earth science and climate change , 2017, Vol.8,Issue 4, pp.1-6 **(Scopus)**. **ISSN: 2157-7617**

[11] Sanobar Chouhan, Saurabh Chodhary, Tarun Upadhaya Ajay Rupani and Pawan Whig "Comparative Study of Various Gates Based in Different Technologies", International Journal Robotics and Automation, Vol 3, Issue 1 , 2017 , pp. 1-7. **ISSN: 0826-8185.(Scopus)**

[12] Pawan Whig and S. N Ahmad," Signal Conditioner Circuit for Water Quality Monitoring Device using Current Differencial Transconductance Amplifier", Information technology and electrical engineering Journal , Vol. 6, Issue 2 , 2017. **ISSN No. : 2306-708X**

[13] Pawan Whig, S N Ahmad and Anupam Priyam," Simulation & performance analysis of various R2R D/A converter using various topologies ", International Journal Robotics and Automation, Vol 2, Issue 1 , 2018 , pp. 128 -131. **ISSN: 0826-8185.(Scopus)**