SMS Spam Detection using Machine Learning using Naïve Bayes

Md. Mihal
Computer Science and Engineering
East West Univerity
Dhaka, Bangladesh
2018-1-60-020@std.ewubd.edu

Md. Muyeen – Ul - Islam
Computer Science and Engineering
East West Univerity
Dhaka, Bangladesh
2018-1-60-022@std.ewubd.edu

Daniel Deep Pereira
Computer Science and Engineering
East West Univerity
Dhaka, Bangladesh
2018-1-60-023@std.ewubd.edu

***Abstract*— Spam messages are a common occurrence in our daily life. Spammers and hackers send spam SMS to hack into our personal information. Spammers steal money and blackmail people using this information. In this research we used multinomial naïve bayes to detect spam words from SMS sent to mobiles. We used a dataset from Kaggle as our source. The detection results of SMS spam words have an accuracy score of 98.2%, precision score of 98.6%, recall score of 99.4%, F1 score of 99%**

Keywords—Spam, spammers, SMS, hack, naïve bayes, multinomial naïve bayes, Kaggle.

I. INTRODUCTION

In this age we are living in, those who use mobile phones face tons of commercial messages incoming on their phones every day. These commercial messages which are sent to the users without their permission are considered spams. Some of these messages are useful to some people but most of them are useless and often irritating. Many of these messages contain malicious links. People who do not understand much about spams click on these links and give away their valuable information to the spammers. Reviews say that almost 60 percent of all the SMSs are spams. Spam SMSs are a severe problem in Bangladesh. In a report of the state-run Bangladesh e-Government Computer Incident Response Team (BGD e-Gov CIRT) it is said that Bangladesh is in the 29th position among the spam sources by country and 7.1 percent regarding the world's spam volume[1]. It has become the biggest cyber threat in Bangladesh. Not only in Bangladesh, these spam SMSs are big problem all across the world. Many attempts have been taken till now to stop these spam messages. Bangladesh Telecommunication Regulator Commission (BTRC) has taken few steps to get rid of the commercial messages sent by the telecommunication companies[2]. However, these are never enough as spammers keep inventing new ways to get through all the barriers. So, people are still practicing different ways to stop these spam messages. If we only look at our personal mobile phones, we can see how severe the problem is as we get more than ten spam messages in the period of only an hour. So, this is no surprise that spam messages are a big concern. People are implementing many approaches and doing many researches to prevent these spam SMSs. In this paper we will discuss about the detection of spam SMSs using Count Vectorizer and Multinomial Naïve Bayes.

II. RELATED WORKS

Spam messages were not a big of a problem not even a decade ago. However, in this age, this problem is getting out of hand. To take care of this problem, different researchers proposed different methods. Dr. Kavita S. Oza have used open source python to detect SMS spam with 98% accuracy [3]. They also used WEKA for preprocessing and studying. Bichitrananda used Support Vector Machine, Decision Tree, K-Nearest Neighbor[4]. Ting S.L used text mining on large datasets using Machine Learning algorithms such as neural network, decision tree, SVM and compared the classifiers depending on computational efficiency and accuracy[5]. Among all the classifiers, Naïve Bayes was the best and efficient.

III. METHOD

In this model we used Multinomial Naïve Bayes and count vectorizer to detect spam words in SMS. Now there are two types of naïve bayes algorithms, one is gaussian and another one is multinomial. Gaussian naïve bayes is the easiest way to work with. Gaussian means normal distribution. In gaussian method only mean and standard deviation are estimated from training data. Gaussian naïve bayes gives the best result when the data is continuous like someone's height and weight, temperature etc.

On the other hand, multinomial naïve bayes is used for categorical text data analysis like word count problems. This is a very popular method for natural language processing. This formula predicts the tag of SMS texts and the calculates the probability of each tag. The output

it produces is the tag with highest probability. This method is very suitable for discrete data. The input it takes is integer word counts. One of its advantages is it can handle big size datasets. In this case the basic naïve bayes formula is

P(A|B) = P(A) * P(B|A)/P(B)

Here the idea is to tokenize all the words and count their frequencies. There are two parts of our category which is spam and ham. Spams are harmful and hams are good messages. For spam the value is 0 and for ham it is 1. The idea is to check the messages and predict whether it is 0 or 1 means spam or ham. Here is a short example of how this works:

Let's say there are total five sentences and from them we consider three sentences for better understanding:

Hello how are you (ham)

Win lottery (spam)

You received money(spam)

As we have used count vectorizer, this will remove the punctuations and stop words and also convert all words into lower cases. Let's say the frequency matrix we will get is like this (for all 5 sentences)

hello	how	win	lottery	received	money	Spam/ham
1	1	0	0	0	0	1
0	0	1	1	0	0	0
0	0	0	0	1	1	0
0	1	1	0	0	0	1
1	0	0	0	1	1	0

The main formula is P(A|B) = P(A) * P(B|A)/P(B)

Here P(A|B) means P (y = spam / ham | sentences)

Now sentences consists of various words, we can consider them like x_1, x_2, $x_3 \ldots x_n$ words. Now when considering the case of ham (y = ham), we can write the formula like this

P (y = ham | (x_1, x_2, $x_3 \ldots$, x_n) =

P(y) * $\prod_{i=1}^{n} P$ (x_1|y = ham) * P (x_2|y = ham) …* P(x_n|y = ham)

Considering the first sentence, x_1 = hello, x_2 = how

When y = ham, probability is 2/5 * 1/2 * 2/2 = 1/10

When y=spam, the equation is

P (y =spam | (x_1, x_2, $x_3 \ldots$, x_n) =

P(y) * $\prod_{i=1}^{n} P$ (x_1|y = spam) * P (x_2|y = spam) …* P(x_n|y = spam)

When y = spam, probability is 3/5 * 1/2 * 0/2 = 0

Naïve bayes takes the greater probability as result. So, it will determine sentence 1 as ham, which in reality is also ham. This will be applicable for all the sentences in dataset. This same process will run our datasets too.

IV. Results and Discussion

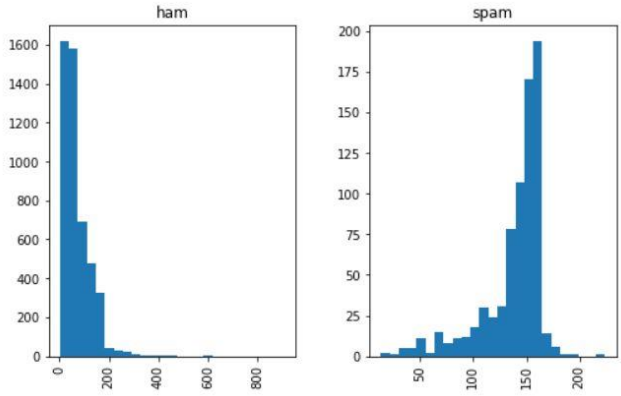
The dataset we used from Kaggle has 5572 rows in it. It had two types of SMSs, hams and spams.

- Preprocessing

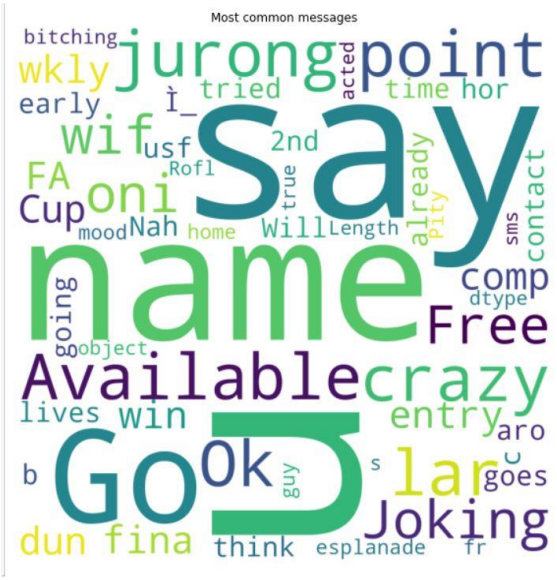
The dataset had to be processed in several manners. First it had some NaN (Not a number) attributes, we dropped those columns as those were unnecessary. Then we renamed remaining two attributes as 'sms' and 'category'. We used count vectorizer as it by default removes the stop words, punctuations and converts letters into lower cases. We used count vectorizer because it takes a word and represent it in a vector of token counts.

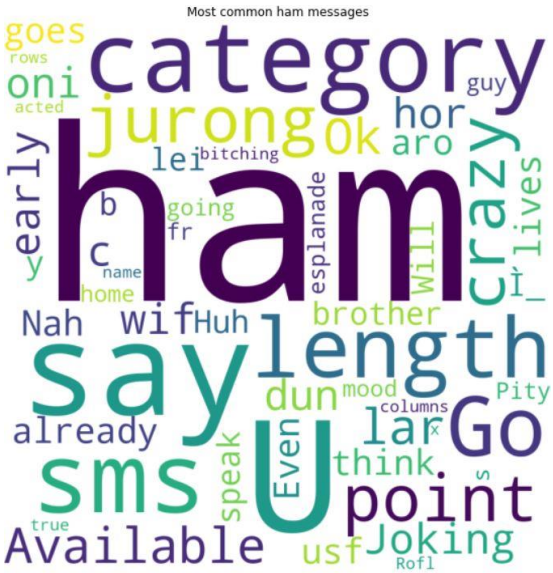
- Testing and Evaluation

As we have applied multinomial naïve bayes in this scenario, the results were pretty good. We separated the ham and spam messages from dataset and plot a histogram for a better view of the dataset.



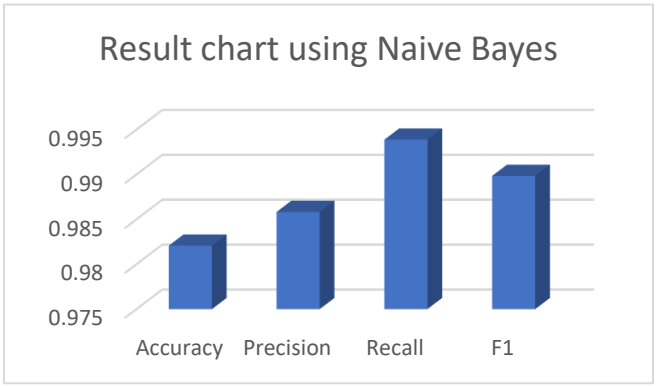
We used word cloud to get a clear view of the most common words in all messages. We also extracted most common ham and spam words too.





After applying naïve bayes, the results are these

Method	Accuracy	Precision	Recall	F1
Multinomial Naïve Bayes	0.98207	0.98582	0.99387	0.98983



V. Conclusion

From the results of using multinomial Naïve Bayes on the dataset about spam SMS, we can say that using Naïve Bayes we got a satisfactory result. The score of accuracy, precision, recall and F1 measure are really high. In the dataset we worked on, it had 5572 rows of data. This dataset can be expanded and more variations can be inserted. This study can be further used for future work and can be more developed in future to get even better score on detecting spam SMSs.

References

[1] "Spam, ransomware Bangladesh's two major cyber threats." https://thefinancialexpress.com.bd/national/spam-ransomware-bangladeshs-two-major-cyber-threats-1609321935 (accessed May 20, 2021).

[2] "How to deal with unwanted SMS | Dhaka Tribune." https://www.dhakatribune.com/opinion/special/2018/01/24/deal-unwanted-sms (accessed May 20, 2021).

[3] "CONTENT-BASED SMS SPAM FILTERING USING MACHINE LEARNING TECHNIQUE - International Journal Of Computer Engineering & Applications." https://www.ijcea.com/content-based-sms-spam-filtering-using-machine-learning-technique/ (accessed May 20, 2021).

[4] "A UGC Approved and Indexed with ICI, DOI, Research Gate, Google Scholar, DPI Digital Library, Scopus (under review), Thomson Reuters (under review)) | Engineering UGC approved journal | computer science UGC approved journal | computer science and enginee." https://www.ijcseonline.org/full_paper_view.ph

p?paper_id=3818 (accessed May 20, 2021).

[5] "(PDF) Is Naïve Bayes a Good Classifier for Document Classification?" https://www.researchgate.net/publication/266463703_Is_Naive_Bayes_a_Good_Classifier_for_Document_Classification (accessed May 20, 2021).