Spam Detection

Jalaj Mehta(202018033)
MSc. Data Science
DA-IICT
Ahmedabad, India
jalajmehta.jm@gmail.com

Anjali Jain(202018036)
MSc. Data Science
DA-IICT
Ahmedabad, India
anjalijain238@gmail.com

Darshan Jain(202018043)
MSc. Data Science
DA-IICT
Ahmedabad, India
darshan.jain159@gmail.com

Aayush Ramrakhyani(202018046)
MSc. Data Science
DA-IICT
Ahmedabad, India
aayushr7777@gmail.com

I.Abstract

Over ongoing years, as the notoriety of cell phone gadgets has expanded, Short Message Service(SMS) has developed into a multi-billion dollar industry. Simultaneously, a decrease in the expense of informing administrations has brought about development in spontaneous business commercials (spams) being shipped off cell phones. Individuals characterize SMS Spam as irritating (32.3%), squandering time(24.8%), also, abusing individual protection (21.3%). In a new report, TrueCaller, the Swedish telephone number ID administration, recognized 8.6 billion spam messages worldwide in 2019. This report additionally recorded the 20 nations that have the most noteworthy number of spam messages per client in a month.

While all nations in the best 3 were in the African mainland followed by Brazil (which gets the generally number of spam brings on the planet) in fourth spot, India wasn't a long ways behind, getting a spot in the best 10 as the eighth-most spam SMS tormented country. Absence of genuine information bases for SMS spam, a short length of messages, and restricted highlights, and their casual language are the elements that may cause the set up email separating calculations to fail to meet expectations in their order.

So in our project, we have made an SMS spam model for detecting the spam using NLTK (Natural Language Toolkit) and we have used many libraries such as pandas,sklearn, seaborn, etc, and even we have checked our model's accuracy, precision, recall f1-score, and support for three different classifiers namely Multinomial Naive Bayes, Random Forest, and Logistic Regression. From which we came to know that the Random Forest classifier is the best and suitable approach for our model.

II. INTRODUCTION

Text informing has been probably the best driver of memberships for cell phones. From the least difficult clamshells to present day cell phones, essentially every cell competent gadget upholds SMS.

Obviously, these frameworks have been focused on widely by spammers. The examination local area has, thusly, reacted with a scope of sifting instruments. Be that as it may, this environment and the messages it conveys have changed significantly in the previous few years. The possibility of this task is to comprehend the mass messages have qualities like spam, including the pervasiveness of a number (like a shortcode or one-time secret word) or a URL, just as a source of inspiration ("click here"), we theorize that

SMS spam channels should change to represent another informing worldview. In this paper, we influence a dataset of almost 400,000 messages gathered throughout 14 months. We get such information by creeping public SMS doors. Clients depend on these public passages to get authentic SMS check messages just as to try not to have their real telephone numbers presented to records that get spam.

III.METHODS

In our project, we have used many tools & libraries, different types of classifiers and at last, we have evaluated classifiers.

1. LIBRARIES :

- Pylab: It is a convenience module that bulks imports for plotting and we have used this to plot the confusion matrix of the classifiers.
- Pandas: In PC programming, pandas is a product library composed for the Python programming language for information control and examination. We have utilized pandas for everything in our model from perusing the record to handling the information and showing the yield
- Seaborn: It is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in the exploration and understanding of data. In our project, we used the seaborn library to generate scatter plots and bar charts.
- String: The Python String Library gives different string Functions, which permits us to perform required procedure on the String information. We have utilized this library in preprocessing like affixing the rundown.
- Re: The Python standard library gives a module to ordinary articulations. Its essential capacity is to offer a pursuit, where it takes a standard articulation and a string.
  NLTK: NLTK represents Natural Language ToolKit. It is a well known library among Python designers who manage Natural Language Processing. NLTK gives the greater part of the capacities needed to handle human language.

  We have utilized NLTK for recognizing spam SMS.

  - In NLTK we have utilized numerous modules for our model, for example,
  - It assists us with disposing of undesirable information and clamor by eliminating accentuations/digits, changing over to bring down cases.
  - The objective of text preprocessing is to change over the content into a structure that is not difficult to measure and examine.
  - Utilizing inbuilt capacity string. accentuation and .isdigit() to check for accentuations and digits and eliminate them.
  - Utilizing re.split() to part message into words(tokens) and utilizing .lower() to change over them into lower case .
  - Stopwords allude to the most usually utilized words in a language.
  - Lemmatization is the cycle where various types of a word are Converted to its root word.
  - Combining tokens is utilized to frame a last book string.
  - Check vectorization includes tallying the quantity of events of each word in a given content.
  - TFIDF(term recurrence reverse record recurrence) is a factual measure that assesses

how important a word is to an archive in an assortment of reports. It reveals to us how significant a word is to a book in a gathering of text. It is determined by duplicating the recurrence of a word, and the opposite report recurrence (how regular a word is, determined by log(number of text/number of text which contains the word)) of the word across a gathering of text.

2. Classifications:

- The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observations into a number of classes or groups. We have used three classifiers in our model namely Multinomial Naïve Bayes, Random Forest, and Logistic Regression.
- Multinomial Naïve Bayes: It is used for integer feature counts.
- Random Forest: It is a group learning strategy for characterization that works by developing a large number of choice trees at preparing time and yielding the class that is the method of the arrangement.
- Logistic Regression:It is a factual model that in its fundamental structure utilizes a calculated capacity to show a double needy variable and we have used it to model the probability of an SMS as spam or not.

3 . Evaluating a Classification model :

Assessing order models A characterization model spots models into one of at least two classifications. For estimating classifier execution, we'll initially present the inconceivably helpful device called the disarray lattice and show how it very well may be utilized to compute numerous significant assessment scores. The principal score we'll examine is exactness.

1. Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right.

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

2. The precision for a class is the number of true positives divided by the total number of elements labeled as belonging to the positive class.

$$\frac{TP}{TP+FP}$$

3. Recall indicates what percentage of the classes we're interested in was actually captured by the model

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

4. F1 score combines precision and recalls relative to a specific positive class -The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

We have also used an in-built function,classification_report() to directly displace the precision, recall,f1-score of the classifiers used in our project.

IV.EXPERIMENTAL RESULTS:

In our Spam Model, we have applied three classifications namely Multinomial Naive Bayes, Random Forest, and Logistic Regression. And we have further compared three of them and checked which one is best and most suitable for our model.

Below are the experimental results:

1. We have found the accuracy of our model by using a confusion matrix and we got different accuracies for different classification such as:

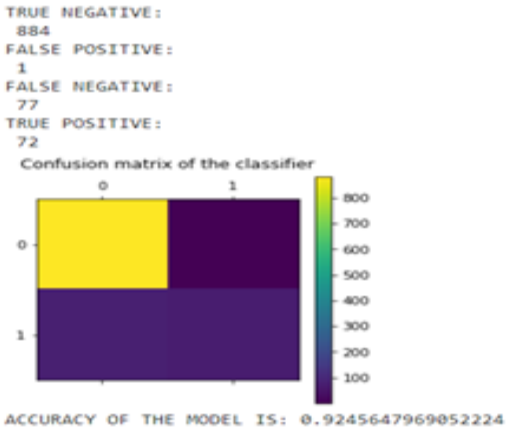
Multinomial Naive Bayes: 0.9246

Random Forest: 0.9816

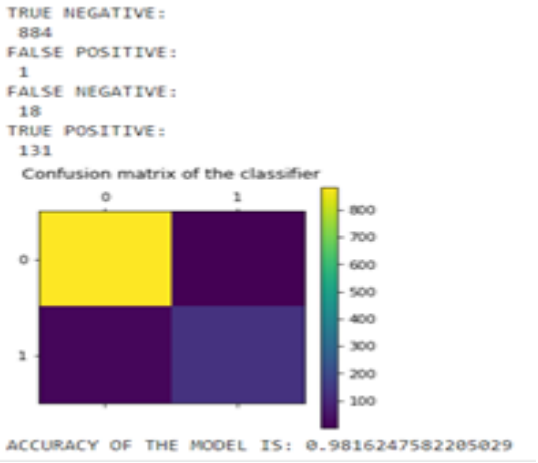
Logistic Regression: 0.9758

So from accuracy, we can say that the random forest classifier is the best and most suitable approach for our spam model.

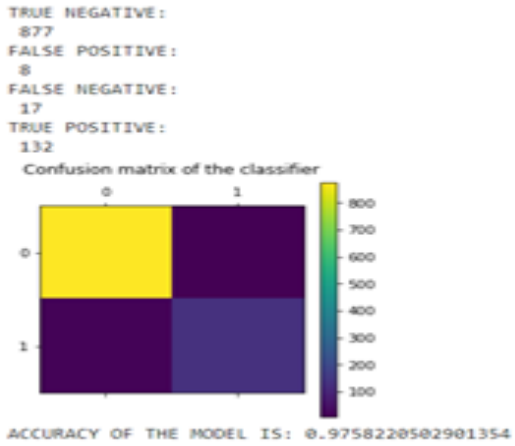
Multinomial Naive Bayes



Random Forest



Logistic Regression



2. We have also found the precision, recall, f1-score, and support of different classifiers for our spam model and compared the results. Below are the respective results of three classifiers:

Multinomial Naive Bayes

	precision	recall	f1-score	support
0	0.92	1.00	0.96	885
1	0.99	0.48	0.65	149
accuracy			0.92	1034
macro avg	0.95	0.74	0.80	1034
weighted avg	0.93	0.92	0.91	1034

Random Forest

	precision	recall	f1-score	support
0	0.98	1.00	0.99	885
1	0.99	0.88	0.93	149
accuracy			0.98	1034
macro avg	0.99	0.94	0.96	1034
weighted avg	0.98	0.98	0.98	1034

Logistic Regression

	precision	recall	f1-score	support
0	0.98	0.99	0.99	885
1	0.94	0.89	0.91	149
accuracy			0.98	1034
macro avg	0.96	0.94	0.95	1034
weighted avg	0.98	0.98	0.98	1034

V.CONCLUSION:

Spam Detection is significant for getting messages and email correspondence. The precise recognition of spam is a major issue and numerous discovery techniques have been proposed by different analysts. Be that as it may, these techniques have an absence of ability to recognize spam precisely and proficiently. The SMS Spam issue is expanding these days with the expansion in the utilization of text informing. Thus, we attempted to take care of this issue by making SMS Spam Detection. The fundamental point of this paper is to contrast diverse AI models with foreseeing whether the message in the dataset is Ham or Spam and predicts the exhibition through precision rule. Utilizing these Models, we can see if the information in the dataset is unsurprising. In this paper, we have made a SMS Spam Detection Model in which we have carried out three order methods dependent on AI calculations to be specific Multinomial Naïve Bayes, Logistic Regression, and Random Forest. What's more, further, we have checked the exactness, accuracy, recall,f1-score, and backing of each of the three classifiers. From the outcomes, we presumed that our model methodology was the correct way as we got every one of the correctnesses above 0.90. Arbitrary Forest Classification Model is the best methodology and generally appropriate for our model as its precision is 0.9816, which is the most elevated.

VI.REFERENCES :

1. Bosaeed, Sahar, Iyad Katib, and Rashid Mehmood. "A Fog-Augmented Machine Learning based SMS Spam Detection and Classification System." In *2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC)*, pp. 325-330. IEEE, 2020.
2. Navaney, Pavas, Gaurav Dubey, and Ajay Rana. "SMS spam filtering using supervised machine learning algorithms." In *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 43-48. IEEE, 2018.
3. Shirani-Mehr, Houshmand. "SMS spam detection using machine learning approach." *unpublished) http://cs229. stanford. edu/proj2013/Shir aniMehr-SMSSpamDetectionUsingMachineLearningApproach. pdf* (2013).GitHubreaves-wisec16.pdf (ufl.edu)

4. Various ways to evaluate a machine learning model's performance | by Kartik Nighania | Towards Data Science
5. 4 Types of Classification Tasks in Machine Learning (machinelearningmastery.com)
6. https://pythonprogramminglanguage.com/logistic-regression-spam-filter/