

SMS Spam Detection Using Machine Learning, Flask and Flutter

Akshay Sable ^{*1}, Nilesh Urkude ^{*1}, Vaibhav Bhade ^{*1}, Shubham Dongare ^{*1},
Chetan Morey ^{*1}

^{*1} Computer Science And Engineering, P.R.Pote College Of Engineering and Management , Amravati Maharashtra, India.

Abstract – ML is termed as Machine Learning; it is the study of computer algorithms which automatically improves through experience. The usage of mobile phones is growing popular in our everyday life. Short Message Service are viewed as most generally applied correspondence administration which is less costly. In any case, this has prompted an increment in cell phones assaults like SMS Spam. Here, Naive bayes algorithm is used in order to differentiate between spam and ham SMS. Spam is the unnecessary fraud messages received whereas ham is legitimate message. The algorithm used here is machine learning classification algorithm, and it is implemented here and can be used in differentiating between spam and ham messages with the help of SMS spam collection data set provided. We train the machine by providing that data set such that it learns from that data and will be able to draw conclusions on its own. Now a days it is very much crucial to identify the spam messages to reduce many frauds happening around the globe. This algorithm can classify with an accuracy of 98.13%.

Keywords: Machine Learning, SMS Spam, Naïve Bayes Classifier, text classification, Sms Spam Detection App, Flutter and Python.

1. INTRODUCTION

Short Message Service is measured as most extensively used message facility. It is a technique of sending short text messages from one device to another. The usage of mobiles is growing everyday as they deliver a huge variation of facilities by dropping the rate of amenities. Due to abundant usage of these services, it has led to growth in mobile device outbreaks like SMS junk. Generally, the word

spam refers to the message which is unsolicited. Simply we can state that spam is a junk text message sent from one device to another in the SMS text format. These spam messages can cause threat to personal data stored in the device.

By the enormous growth in population and increase of all these technological aspects have been growing extremely which in turn increases unanimous spreading of such threat due to less effective security control measure and in order to solve such problems many researchers have developed many techniques to solve and protect the devices from such threats in many different ways. The main motive is to provide privacy, convenience and harmony. The classifier used to build these models are Support Vector Machine and

Naïve Bayes Classifier, these the two mostly used traditional classifiers. And by convolutional Neural Networks also we can achieve the required model.

Essentially while training the model firstly, we want to consider a data set, here SMS Spam Collection dataset is used and it is divided into train and test data set and the Naïve Bayes Classifier is used in training and evaluation of the model.

2. LITERATURE SURVEY

A. LOGISTIC REGRESSION

This technique that denotes binary values (0 or 1, true or false, yes or no), implying that the results can only be in two forms. Finding the probability of a favorable or failed event can be seen as an example of binary values.

Limitations and disadvantages:

-It constructs linear boundary.

-Discrete no. of set.

-The assumption of linearity between dependent and independent variable.

B. SUPPORT VECTOR MACHINE

The Support Vector Machine (SVM) is a Supervised Learning algorithm, the Support Vector model is used for classification problems.

Limitations and disadvantages:

-Not suitable for large data set.

-No. of features for each data point exceeds it will under-perform.

C. NEURAL NETWORK

Also known as ANN. Neural networks are potent algorithm for solving any machine-learning problem that requires classification

Limitations and disadvantages:

-Assurance of proper network structure

-The difficulty of showing the problem to the network

The duration of the network is unknown

D. BOOSTING CLASSIFIER

Also known as Adaboost. The boosting classifier is based

on the reassessing the weight of weak classifiers. The error term will be calculated and re-weight will be assigned which will strengthen the accuracy of classifiers.

Limitations and disadvantages:

- Time consuming.
- Hard to implement, Complex.

E. ENSEMBLE CLASSIFIER

Used to provide a better prediction. the problem of concept drift is also solved by it. Limitations and disadvantages:

- Difficult to interpret, can be confusing.
- Can't help unknown difference.

F. DECISION TREE

A decision tree is a flow chart like construction, where. Internal node or non- leaf node= Test on attribute Branch = shows outcome of the test Leaf node= holds a class label Top node is called root node.

Limitations and disadvantages:

- Unstable, Often inaccurate.

G. NAIVE BAYES

It is an algorithm of Machine Learning which comes under classification technique. It's a Supervised technique that uses the Bayes theorem of probability to predict the class of enigmatic datasets. In simple terms, a Naive Bayes algorithm anticipates that the presence of one item in a class is unrelated to the presence of another. Naive Bayes model is not difficult to assemble and especially valuable for enormous informational indexes. It works on the probability principle called and bayes theorem which is shown in fig.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Probability of B occurring given evidence A has already occurred (points to $P(B|A)$)
 Probability of A occurring (points to $P(A)$)
 Probability of A occurring given evidence B has already occurred (points to $P(A|B)$)
 Probability of B occurring (points to $P(B)$)

Fig :- Bayes Theorem

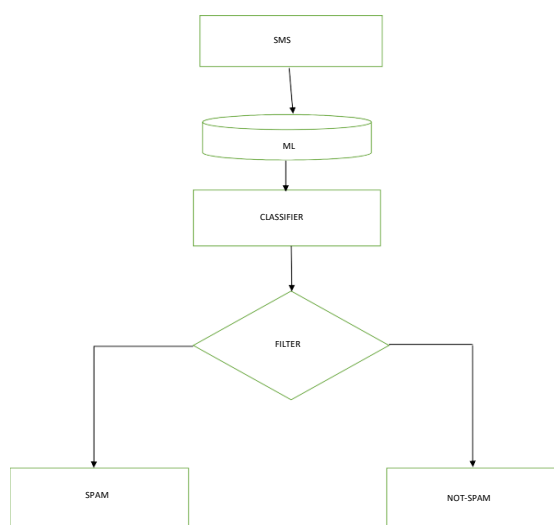
H. K- NEAREST NEIGHBOUR

K-nearest neighbours is a supervised classification algorithm. This algorithm has some data point and data vector that are separated into several classes to predict the classification of new sample point.

Limitations and disadvantages:

- It is called as lazy algorithm means it tries to only memorize the process.
- It doesn't learn on its own.
- It doesn't take its own decision.

3. WORKPLAN



4. METHOD OF PROPOSED MODEL

A. DATA COLLECTION

Collecting dataset for training the model/classifier. To train a model/ classifier we required data that helps the algorithm in learning the patterns of spam.

B. DATA CLEANING

In order to train the supervised machine learning model the prior thing to do is collect the dataset. Here the data set considered is sms spam collection which is taken from UCI machine learning repository. It contains the collected sms with label whether it is spam or ham message. The glimpse of data is as shown in the fig.

| label | message | Length |
|-------|--|--------|
| 0 | ham Go until jurong point, crazy.. Available only ... | 111 |
| 1 | ham OK lar... Joking wif u oni... | 29 |
| 2 | spam Free entry in 2 a wkly comp to win FA Cup fina... | 155 |
| 3 | ham U dun say so early hor... U c already then say... | 49 |
| 4 | ham Nah I don't think he goes to usf, he lives aro... | 61 |

Fig:- head of dataset

C. DATA ANALYSING

EDA (Exploratory Data Analysis) is unavoidable and one of the major step to fine-tune the given data set(s) in a different form of analysis to understand the insights of the key characteristics of various entities of the data set like column(s), row(s) by applying Pandas, NumPy, Statistical Methods, and Data visualization packages.

D. PROCESSING TEXT

Text processing is automation of analysing electronic text. This allows machine learning models to get structured information about the text to use for analysis, manipulation of the text, or to generate new text.

E. BUILDING AND TRAINING MODEL

Naïve Bayes Algorithm/ Classifier: Naive Bayes algorithm exemplifies a supervised learning technique and at the same time a statistical technique for classification. It acts as a fundamental probabilistic model and let us seize ambiguity about the model in an ethical way by influencing the probabilities of the results. It is used to provide solution to analytical and predictive problems.

Bayes' Theorem finds the probability of an event

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation

where A and B are events and $P(B) \neq 0$. $P(A)$ is the priori of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B). $P(A|B)$ is a posteriori probability of B, i.e. probability of event after evidence is seen

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Now, with regards to our dataset, we can apply Bayes' theorem in following way:

where, y is class variable and X is a dependent feature vector (of size n) where:

$$X = (x_1, x_2, x_3, \dots, x_n)$$

5. RESULTS AND DISCUSSION

In this final step, on our prepared dataset, we will test our classification model and also measure the efficiency of SMS spam detection on our dataset. To assess the efficiency of Our defined category and make it comparable to existing approaches .SMS Spam detectors are beneficial and used to future enhancement as this will detect the spam messages and network resources many upcoming detectors are upcoming in future enhancement.

Once you have done all of the above, you can start running the API by either double click temp.py or executing the command from the Terminal and open Application so the output will be in following:

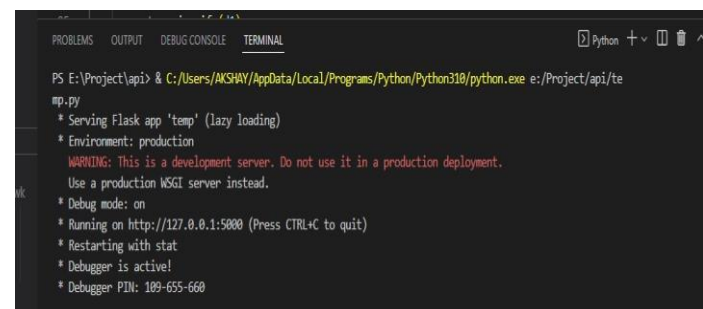


Fig :- img command exe

Presently you could open an Application check Messages.

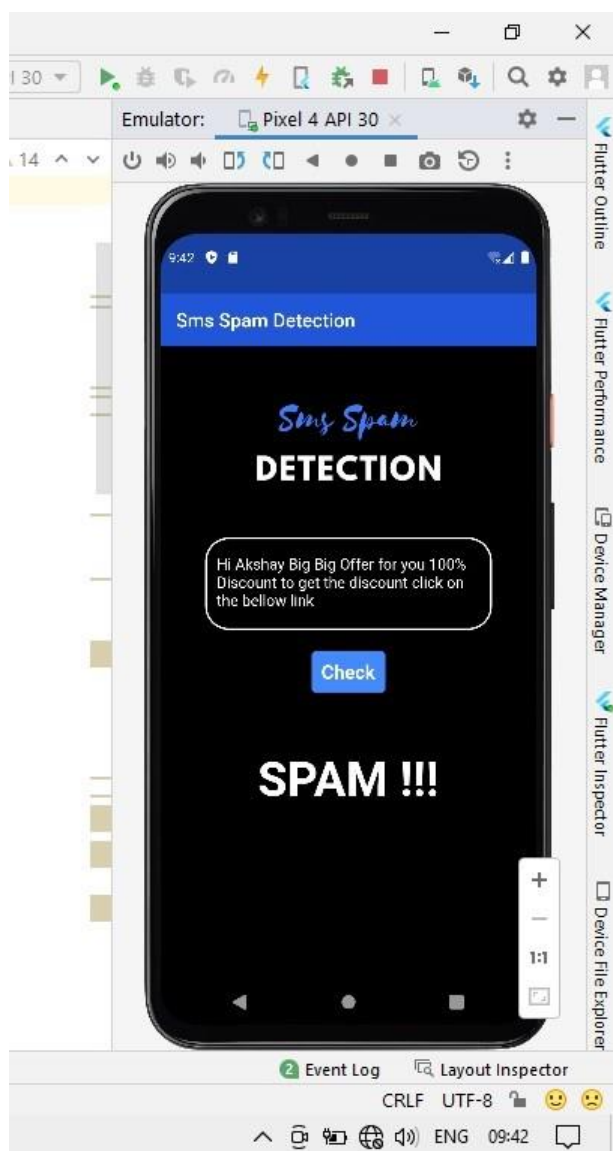


Fig :- shows spam Message which is generally sent in Systems

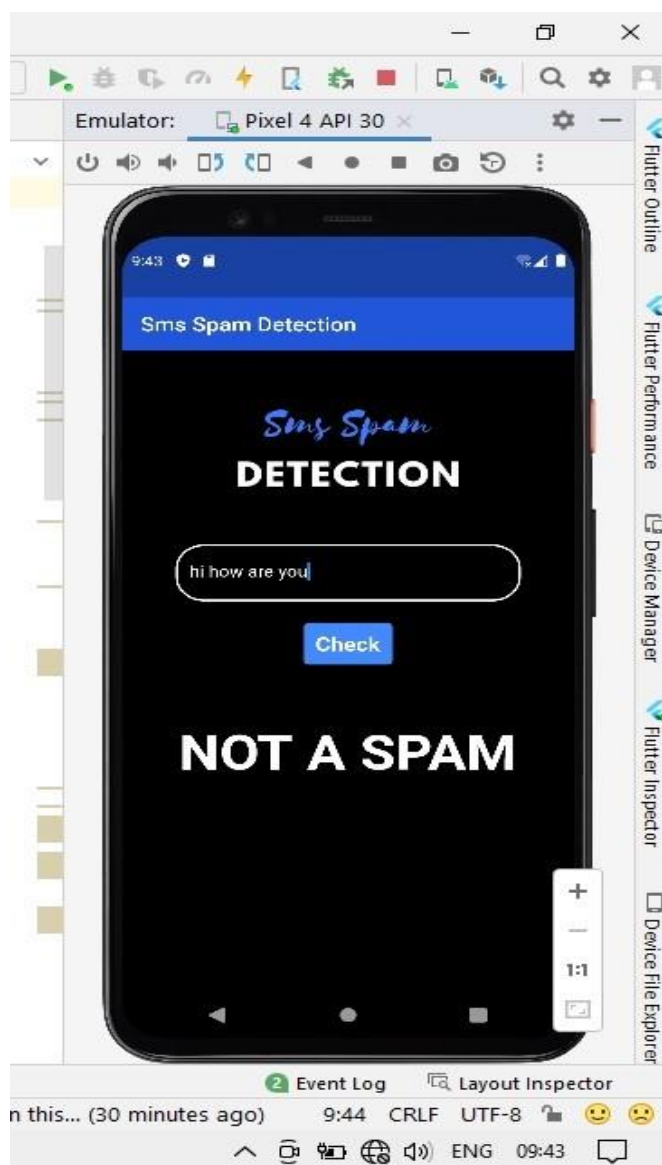


Fig :- shows Not a Spam message

6. CONCLUSION

In this paper various algorithms are proposed and compared. After comparing and referring Naïve Bayes algorithm, is the best algorithm with high accuracy, less time to build the model and testing is also better compared to other algorithms. Said algorithm is used in email spam detection.

REFERENCES

1. V. Christina, S. Karpagavalli, G. Suganya "Email Spam Filtering using Supervised Machine Learning Techniques" Vol. 02, No. 09, 2010, 3126-3129.
2. 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
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. 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.
5. I. Murynets and R. P. Jover, "Analysis of SMS Spam in Mobility Networks," International Journal of Advanced Computer Science, vol. 3, 2013.