International Journal of Computing, Programming and Database Management

E-ISSN: 2707-6644
P-ISSN: 2707-6636
IJCPDM 2020; 1(2): 38-43
Received: 20-05-2020
Accepted: 24-06-2020

Venkatalakshmi Katta
GATE College, Tirupati,
Andhra Pradesh, India

Corresponding Author:
Venkatalakshmi Katta
GATE College, Tirupati,
Andhra Pradesh, India

SMS classification using machine learning algorithms for the risky messages identifying

Venkatalakshmi Katta

DOI: https://doi.org/10.33545/27076636.2020.v1.i2a.17

Abstract

The bulk of SMS that the Quick Response Team and Rescue Agencies received during disasters made it hard for them to categorize responses based on priorities. This paper provides a method that classifies SMS received by the agency as Spam, Invalid, Alert 1 Alert 2, and Alert 3. This method allows proper response to be extended to those asking for it based on prevailing needs. This also provides a chance to ignore insignificant messages and save precious time that may be incurred by merely dealing with unimportant messages. The implementation of Naïve Bayes Algorithm, a self-learning algorithm, and together with Natural Language Processing was utilized in this research. Extension of the method is however devised in order to cover the irregularity of the data to process. Test results of the classification method showed success in its implementation and since it is a self-learning process, the method gets better and became more accurate through time.

Keywords: Naive Bayes formula, messages, prediction, training and testing

Introduction

Mining information from natural languages sent through text message using mobile devices could be of significance during a disaster and crisis management. With the creation of the different government agencies [3] to respond and mitigate crises and disasters brought by climate change, risk reduction and response became accessible and readily available. However, these agencies rely heavily on only information provided and also through requests for their intervention during disasters. It would be of great help to the agency if they can initiate the intervention from their position. With the introduction of wireless communication, where almost everybody from all walks of life is hooked to this technology, the posts, messages, and conversations that they contribute to this technology are filled with unharnessed information [1, 4]. If proper technology is utilized, the unstructured information in the form of SMS could be of something significant [5]. However, in-depth processes using technology must be devised in order to turn this information into something useful.

Related works

Performance Analysis of Naive Bayes and J48Classification Algorithm

Classification is a significant information mining method with a wide scope of uses to arrange the various kinds of information utilized in pretty much every field of our lives. Classification is utilized to characterize a gathering of predefined classes as indicated by the traits of the thing. This paper reveals an insight into execution assessment dependent on good and bad situations of information classification utilizing the Naïve Bayes and J48 classification calculation. The Naive Bayes calculation depends on likelihood and the j48 calculation depends on the choice tree. This paper embarks to make a near assessment of the classifiers NAIVE BAYES and J48 with regards to the Bank Dataset to build the genuine positive rate and diminish the bogus positive pace of defaulters, not just accomplishing high classification exactness utilizing the WEKA instrument.

Influence of Word Normalization on Text Classification: In this paper we focus on the comparison of various lemmatization and steaming algorithms that are often used in nature language processing (NLP). We describe the algorithm in detail and compare it with other widely used algorithms for word normalization in two different entities. We provide better results obtained by our EWN-based lemmatization approach compared to other methods. We also discuss the impact of the term generalization on classification work in general.

Real-time Crisis Mapping of Natural Disasters Using Social Media

The proposed online networking emergency planning stage for catastrophic events utilizes areas from Gazetteer, Street Map and Voluntary Geographic Information (VGI) hotspots for calamity chance regions and geologically coordinates ongoing tweet information streams. The creators utilize factual investigation to make constant emergency maps. Geoparsing results can be benchmarked against effectively distributed works and assessed in multilingual datasets. Two contextual analyses contrast five-day tweet emergency guides and authority post-occasion sway evaluation from the US National Geospatial Agency (NGA), arranged from checked satellite and elevated picture sources.

Location Identification for Crime & Disaster Events by Geoparsing Twitter

Geoparsing means automatically identifying locations in a text. Positioning in messages during crime and catastrophic events is important because they help to quickly locate the place where dispatchers can dispatch aid.

The use of social media during such crisis events is rapidly increasing worldwide and in India. We treat Twitter as the source of the message because it is real-time, visual and can handle large amounts of data.

We collect tweets in real time and apply those tweets to crisis situation and location information. Capturing location information to the level of streets & buildings helps determine the exact location of the event; this is done with the help of NLP methods. We use classifiers to classify tweets to get the event to occur.

3. Proposed System

The Bayesian classifier in this research is used to represent the probability distribution of identified text messages of SMS. This method was formulated to solve the classification problem associated with this research. Considering the independence of the tokens among each other and also with respect to the output, the Naïve Bayesian Classifier is the most appropriate method that can efficiently do the job.

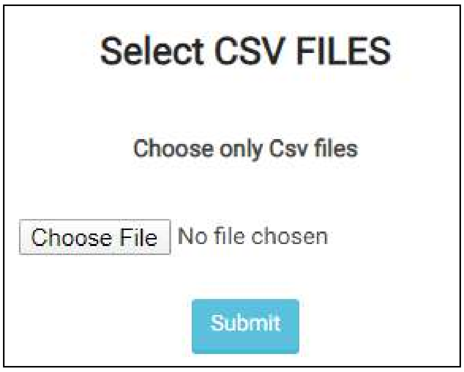
If the input is considered to be a set of attributes, like words, then using a Bayesian network, we can calculate the probability of whether a message belongs to a specific class or not.

The process is composed of two phases: the training phase and the classification phase. In the training phase, the filter is trained using a known collection of words for classified messages. A database of tokens appearing in each corpus and their total occurrences are maintained in a database. Based on their occurrences in each set of classified messages, each token is assigned a probability for its capacity to determine a message of its classification.

Naive Bayes

The problem of classification predictive modelling can be framed as calculating the conditional probability of a class label given a data sample. Bayes Theorem provides a principled way for calculating this conditional probability, although in practice requires an enormous number of samples (very large-sized dataset) and is computationally expensive.

4. Results and Discussions



Select CSV FILES

Choose only Csv files

Fig 1: Uploading CSV File

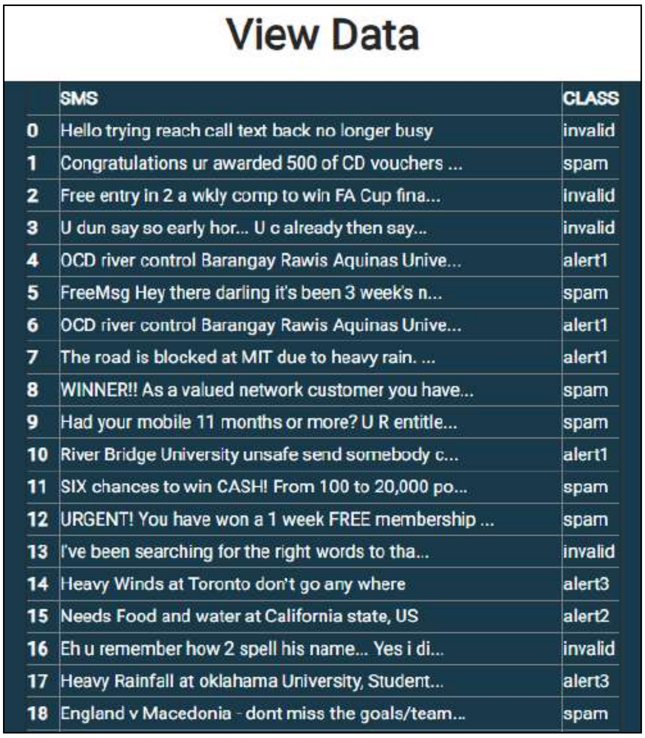
Here we are uploading the sample data set for further process. Training and testing by using a specific model.

Select CSV FILES

Choose only Csv files

Fig 2: CSV File Upload Success

After uploading the dataset, the system shows the successful message (Data uploaded successfully).

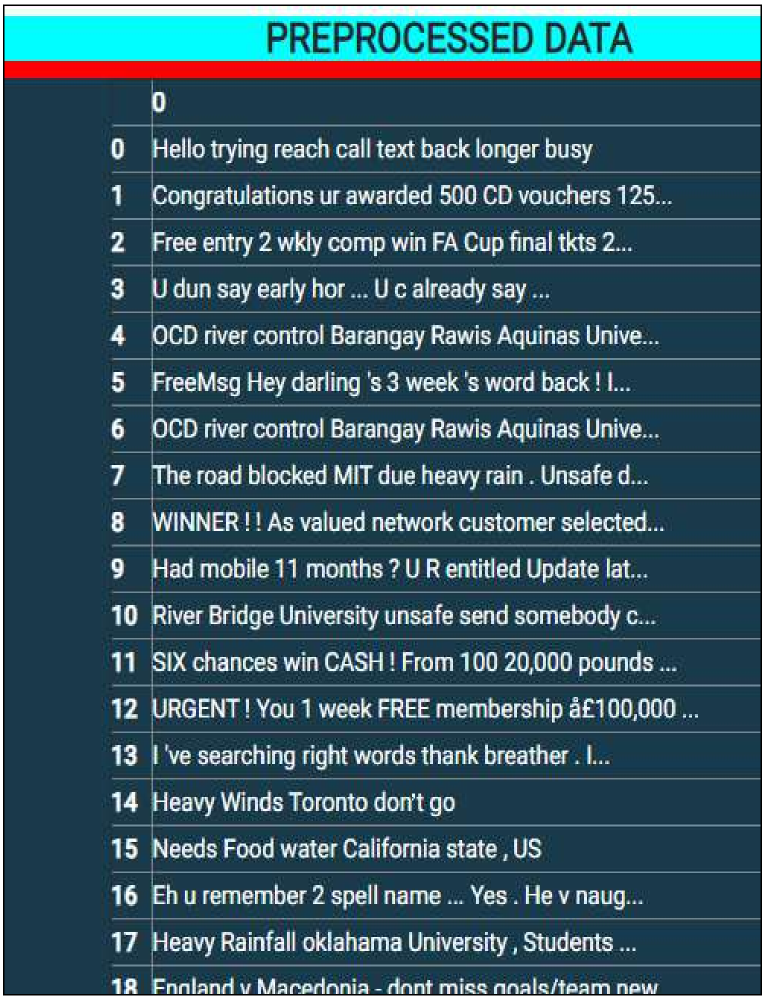


View Data

	SMS	CLASS
0	Hello trying reach call text back no longer busy	invalid
1	Congratulations ur awarded 500 of CD vouchers ...	spam
2	Free entry in 2 a wkly comp to win FA Cup fina...	invalid
3	U dun say so early hor... U c already then say...	invalid
4	OCD river control Barangay Rawis Aquinas Unive...	alert1
5	FreeMsg Hey there darling it's been 3 week's n...	spam
6	OCD river control Barangay Rawis Aquinas Unive...	alert1
7	The road is blocked at MIT due to heavy rain. ...	alert1
8	WINNER!! As a valued network customer you have...	spam
9	Had your mobile 11 months or more? U R entitle...	spam
10	River Bridge University unsafe send somebody c...	alert1
11	SIX chances to win CASH! From 100 to 20,000 po...	spam
12	URGENT! You have won a 1 week FREE membership ...	spam
13	I've been searching for the right words to tha...	invalid
14	Heavy Winds at Toronto don't go any where	alert3
15	Needs Food and water at California state, US	alert2
16	Eh u remember how 2 spell his name... Yes i di...	invalid
17	Heavy Rainfall at oklahama University, Student...	alert3
18	England v Macedonia - dont miss the goals/team...	spam

Fig 3: View Data

In Fig 3 we can view the uploaded sample data set like in a data frame manner.



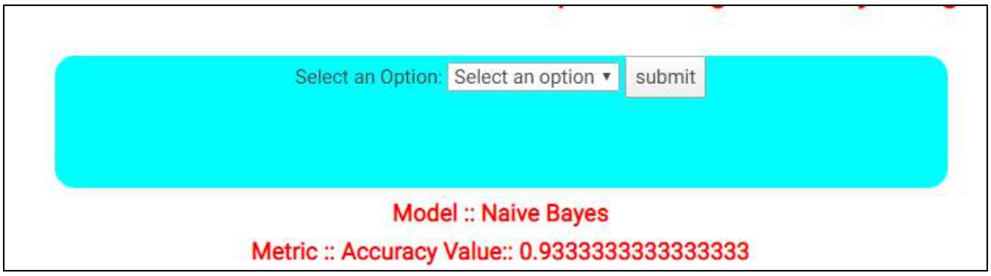
PREPROCESSED DATA

	0
0	Hello trying reach call text back longer busy
1	Congratulations ur awarded 500 CD vouchers 125...
2	Free entry 2 wkly comp win FA Cup final tkts 2...
3	U dun say early hor ... U c already say ...
4	OCD river control Barangay Rawis Aquinas Unive...
5	FreeMsg Hey darling 's 3 week 's word back ! I...
6	OCD river control Barangay Rawis Aquinas Unive...
7	The road blocked MIT due heavy rain . Unsafe d...
8	WINNER ! ! As valued network customer selected...
9	Had mobile 11 months ? U R entitled Update lat...
10	River Bridge University unsafe send somebody c...
11	SIX chances win CASH ! From 100 20,000 pounds ...
12	URGENT ! You 1 week FREE membership â£100,000 ...
13	I 've searching right words thank breather . I...
14	Heavy Winds Toronto don't go
15	Needs Food water California state , US
16	Eh u remember 2 spell name ... Yes . He v naug...
17	Heavy Rainfall oklahama University , Students ...
18	England v Macedonia - dont miss goals/team new

Fig 4: Preprocessed Data

In the above-mentioned Fig 4 we are applying the Preprocessing operation on our sample data set.

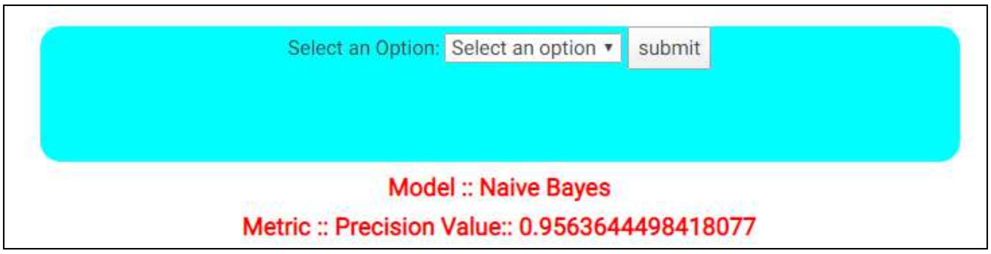
Model Training
Performance Metrics



Model :: Naive Bayes

Metric :: Accuracy Value:: 0.9333333333333333

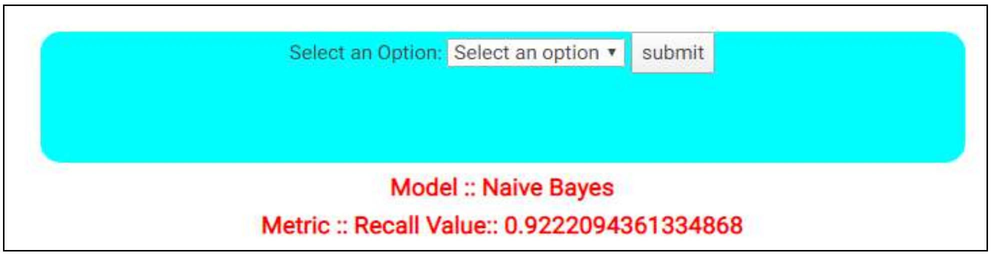
Fig 5: Accuracy



Model :: Naive Bayes

Metric :: Precision Value:: 0.9563644498418077

Fig 6: Precision



Model :: Naive Bayes

Metric :: Recall Value:: 0.9222094361334868

Fig 7: Recall

In Fig 5, 6, 7 are the Results of some matrices like Accuracy, Precision, and Recall for implemented machine learning algorithm called as Navie Bayes Algorithm.

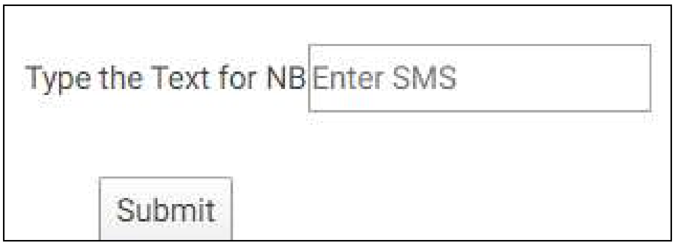
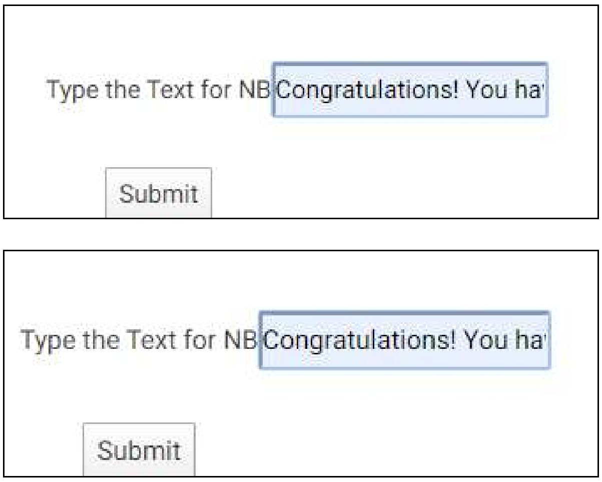


Fig 8: Prediction

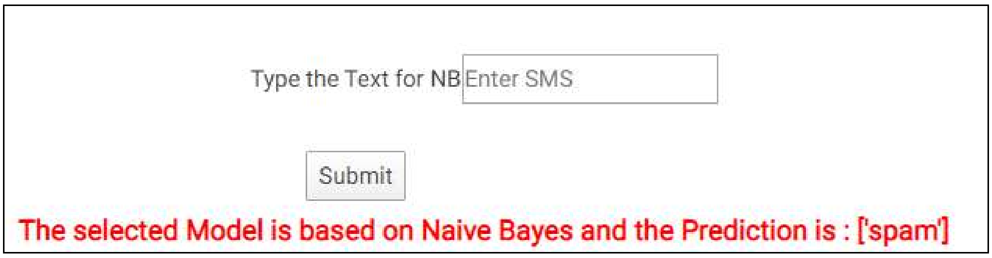
Here in the above Fig 8 we are checking the prediction of the given sample data set. After prediction we are predicting the result for given input that the message is spam or ham.



Congratulations! You ha

Congratulations! You ha

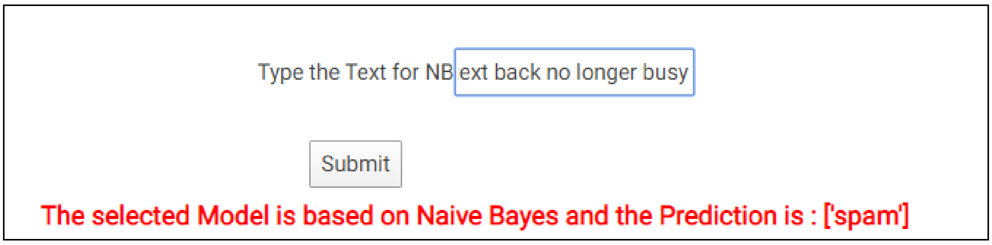
Fig 9: Spam Message



The selected Model is based on Naive Bayes and the Prediction is : ['spam']

Fig 10: Spam Correctly Predicted

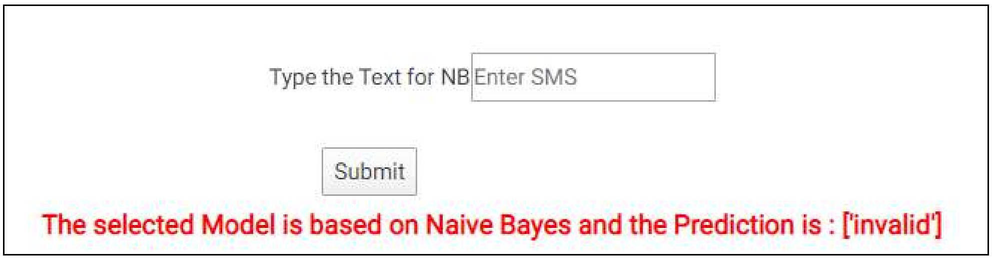
In the above Fig 9 and 10 we are predicting that the given input message is spam or ham here our model predicting the given input message is spam and it is exact outcome of the applied model.



ext back no longer busy

The selected Model is based on Naive Bayes and the Prediction is : ['spam']

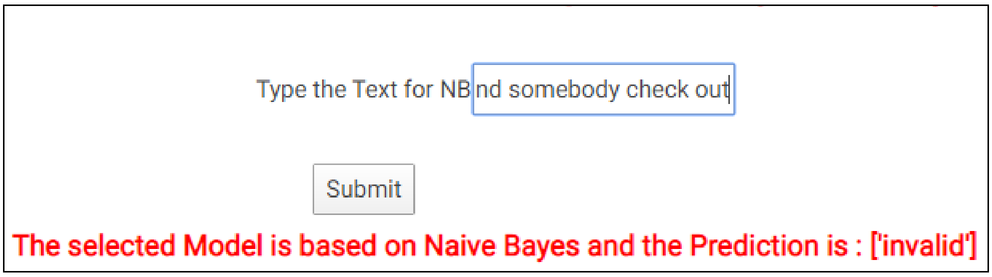
Fig 11: Invalid Message



The selected Model is based on Naive Bayes and the Prediction is : ['invalid']

Fig 12: Invalid Message correctly predicted

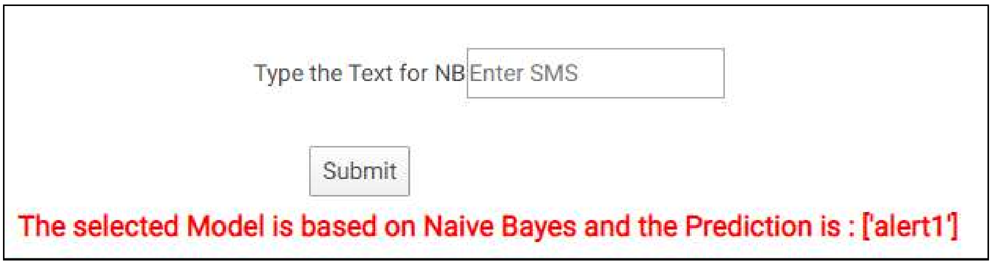
Here in the above Fig 11 and 12 we are predicting that the given input message is invalid and that is correctly predicted by our applied model.



Type the Text for NB nd somebody check out

The selected Model is based on Naive Bayes and the Prediction is : ['invalid']

Fig 13: Alert1 Message

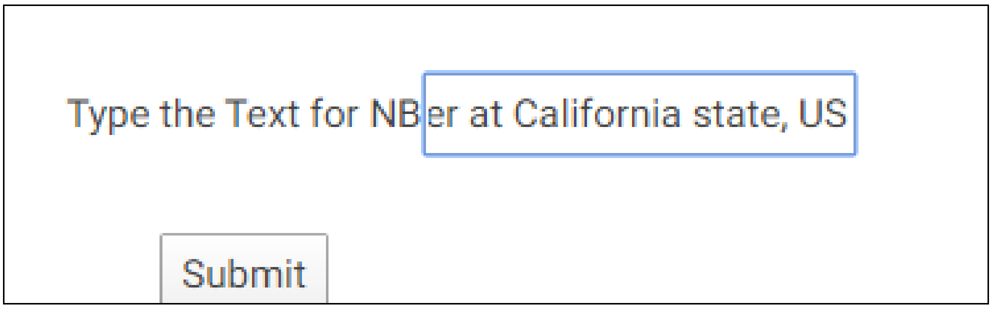


Type the Text for NB Enter SMS

The selected Model is based on Naive Bayes and the Prediction is : ['alert1']

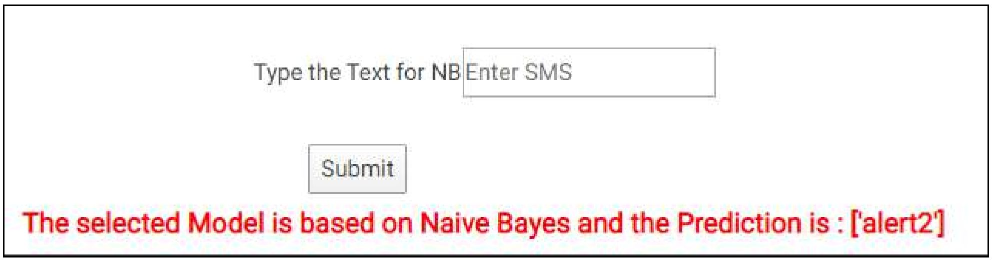
Fig 14: Alert1 Message correctly predicted

Here in the above shown Fig 13 and 14 we are getting output from the applied model as the given input is an alert message and the applied model is also predicting the result as exactly it is alert message.



Type the Text for NB er at California state, US

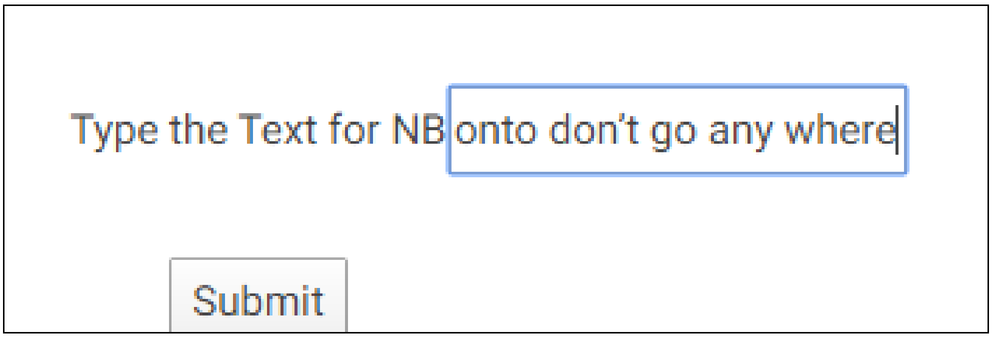
Fig 15: Alert 2 Message



Type the Text for NB Enter SMS

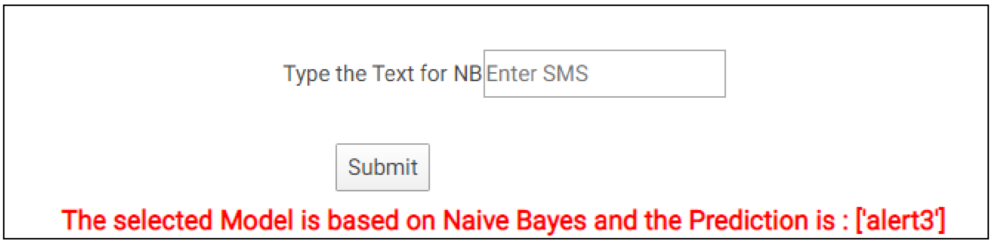
The selected Model is based on Naive Bayes and the Prediction is : ['alert2']

Fig 16: Alert 2 Message correctly predicted



Type the Text for NB onto don't go any where

Fig 17: Alert 3 Message


Type the Text for NB Enter SMS

Submit

The selected Model is based on Naive Bayes and the Prediction is : ['alert3']

Fig 18: Alert 3 Message correctly predicted

In the above-mentioned Fig 15, 16, 17 and 18 predicting the exact outputs for given inputs, the given input message is alert messages.

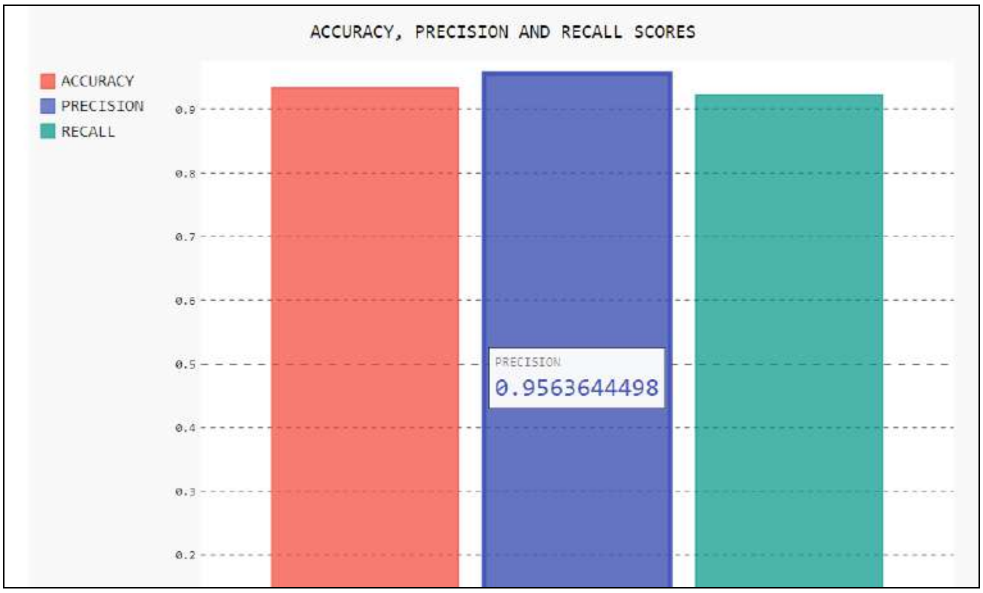


Fig 19: Performance Metrics Graph

In Fig 19 accuracy, precision, recall scores are shown for given input data. Accuracy is preferred for balanced data and precision and recall are preferred for unbalanced data.

Conclusion

This study proved the capability of the extended Naïve Bayes Formula to classify SMS messages according to five different classifications based on the collection of pre-classified information used as the learned classifier. The study generally correctly classified test data according to specific classifications up to 89% accuracy. The 11% that falls within the False-Negative result is attributed to the number of entries in the dataset that served as the learned classifier of the Naïve Bayes Algorithm. Clearly, there is a necessity to increase the number of pre-classified entries in the learned classifier to further improve the capability of the method to correctly classify SMS inputs. However, since the method is self-learning, it recommended that the system be used in actual operation to meet the required high accuracy classification capability of the process.

References

1. Tina R Patil *et al*. Performance Analysis of Naive Bayes and J48Classification Algorithm for Data Classification. International Journal of Computer Science and Applications. 2013, 6(2). ISSN: 0974-1011 (Open Access)
2. Sudarsan VS, Govind Kuma. Voice call analytics using natural language processing. Int J Stat Appl Math 2019;4(6):133-136.
3. Office of the Civil Defense Website, Mandate, Mission, Vision, and Objectives http://ocd.gov.ph/index.php/about-ocd/mandate-missionand-vision
4. Stuart E Middleton, Lee Middleton, Stefano Modafferi, Real-time Crisis Mapping of Natural Disasters Using Social Media, University of Southampton IT Innovation Centre, 2014.
5. Nikhil Dhavase. Location Identification for Crime & Disaster Events by Geoparsing Twitter, 2014, (M.E. Student) Dept. of Information Technology Pune Institute of Computer Technology Pune, India, International Conference for Convergence of Technology, 2014, 978-1-4799-3759-2/14/$31.00©2014 IEEE
6. Michal Toman *et al*. Influence of Word Normalization on Text Classification, University of West Bohemia, Faculty of Applied Sciences, Plzen, Czech Republic.