INTERNATIONAL JOURNAL OF RESEARCH AND DEVELOPMENT IN ENGINEERING SCIENCES
e-ISSN: 2582 - 4201 Volume: 2 | Issue: 6 | Year: 2020

Architectural approach of Machine Learning Algorithms NAVIE BAYES and Decision Trees for Disaster Management

Dr. L.PRASANNA KUMAR
Professor, Department of Computer Science and Engineering,
Dadi Institute of Engineering and Technology (DIET), Anakapalle
Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh

N.SIREESHA
M.Tech Student, Department of Computer Science and Engineering,
Dadi Institute of Engineering and Technology (DIET), Anakapalle
Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh

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.cen*

In recent years the highest degree of communication happens through e-mails which are often affected by passive or active attacks. Effective spam filtering measures are the timely requirement to handle such attacks.

Keyword: *Machine Learning, NAVIE BAYES, Decision Trees, Disaster Management*

1. INTRODUCTION

Maximum number of industries and organizations depends on the usage of machines, electronic widgets, and various recourses for completing different tasks. 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.

This study is focused on classifying SMS information sent through mobile devices that are coursed through authorized agencies to determine the authenticity of the message and its prevailing needs. The primary intention is to provide relevant and correct action towards a valid request and to set aside spam and insignificant messages as well. This was made possible through the use of basic NLP processes and Naïve Bayes Algorithm. The expected system has the capability to classify SMS messages as Spam, Invalid, Alert 1 (does not need immediate attention), Alert 2 (requires attention within the day) and Alert 3 (requires immediate attention ASAP).

2. RELATED WORK

There has been numerous numbers of studies on active learning for text classification using machine learning techniques [9]-[11], probabilistic models [12], [13]. The query by committee algorithm (Seung et al. 1992, Freund et al., 1997) used priori distribution than hypothesis. The popular techniques for text classifications [11]-[14], nearest neighbors and later on Support Vector Machine [12]. Though there is lot of techniques and algorithms which have been proposed so far, the text classification is not yet accurate and faultless and still in demand of improvement. Two types of SMS classification exists in the current mobile phones and they are enlisted as Black and White [14]. These kinds of techniques are based on the previously known keywords and patterns. These techniques are currently available to the numerous number of cell phone operating

e-ISSN: 2582 - 4201 Volume: 2 | Issue: 6 | Year: 2020

systems. These techniques are also recalled as Spam SMS blocker in Google android phones and SMS spam runner in Symbian Operating Systems.

The Key solution is to extend the class conditional probability estimation in the Bayes model to handle pdf. In the project [3], the author has addressed the problem of extending traditional naïve Bayes model to the classification of uncertain data. The key problem in naive Bayes model is a class conditional probability estimation, and kernel density estimation is a common way for that. The kernel density estimation method has extended to handle uncertain data. This decreases the problem to the consideration of double-integrals. For specific kernel functions and possibility distributions, the double integral can be analytically assessed to give a closed-form formula, authorizing an efficient formula-based algorithm.

In general, however, the double integral cannot be simplified in closed forms. In this case, a sample-based approach is proposed. Extensive experiments on several UCI datasets show that the uncertain naive Bayes model considering the full pdf information of uncertain data can produce classifiers with higher accuracy than the traditional model that utilizing the mean as the representative value of uncertain data.

Time complexity analysis and performance analysis based on experiments perform that the formula-based approach has special advantages over the sample-based approach. The naive Bayes classifier especially simplifies learning by expecting that features are autonomic given class. However, independence is generally a poor expectation, in practice, naive Bayes sometimes approaches well with more sophisticated classifiers. [4] Despite its unrealistic independence assumption, the naïve Bayes classifier is surprisingly effective in practice since its classification decision may often be correct even if its probability estimates are inaccurate. Although some optimality conditions of naive Bayes have been already identified in the past, a deeper understanding of data Characteristics that affect the performance of naive Bayes is still required. Instead, a better predictor of accuracy is the information loss that features contain the class when assuming naive Bayes model.

3. METHODOLOGY

NAIVE BAYES & DECISION TREES

3.1. NAIVE BAYES

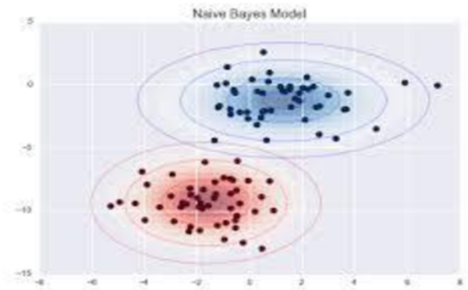
The study of Patil [1] which is focused on the performance analysis of Naïve Bayes and J48 classification algorithm showed an accuracy of both in classifying specimens. Naïve Bayes being probabilistic in nature is more appropriate to use in this study. The paper [2] which is focused on normalization of text provided a pre-processing technique that this study considered employing especially in data cleansing and preparation prior to the text classification process. Stop words are first removed prior to normalization which converts words into its basic forms. Mobile SMS which serves as the input of the operation goes through a process that sanitizes the input from unnecessary words and ambiguous terms. The data undergo stop-words removal and normalization. The sanitized data will serve as the input to the next process which is the actual classification process. The result is the classified text message..

3.1.2. Naive Bayes algorithm

Naive Bayes is a probabilistic algorithm based on the Bayes Theorem used for email spam filtering in data analytics. If you have an email account, we are sure that you have seen emails being categorized into different buckets and automatically being marked important, spam, promotions, etc. Isn't it wonderful to see machines being so smart and doing the work for you? More often than not, these labels added by the system are right. So does this mean our email software is reading through every communication and now understands what you as a user would have done? Absolutely right!

In this age and time of data analytics & machine learning, automated filtering of emails happens via algorithms like Naive Bayes Classifier, which apply the basic Bayes Theorem on the data. In this article, we will understand briefly about the Naive Bayes Algorithm before we get our hands dirty and analyze a real email dataset in Python. This blog is second in the series to understand the Naive Bayes Algorithm. You can read part 1 here in the introduction to Bayes Theorem & Naive Bayes Algorithm blog.

INTERNATIONAL JOURNAL OF RESEARCH AND DEVELOPMENT IN ENGINEERING SCIENCES
e-ISSN: 2582 - 4201 Volume: 2 | Issue: 6 | Year: 2020



The Naive Bayes Classifier Formula

One of the simplest yet powerful classifier algorithms, Naive Bayes is based on Bayes' Theorem Formula with an assumption of independence among predictors. Given a Hypothesis A and evidence B, Bayes' Theorem calculator states that the relationship between the probability of Hypothesis before getting the evidence P(A) and the probability of the hypothesis after getting the evidence P(A|B) is:

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

3.1.3 Decision Tree Classification Algorithm

a) Decision Tree is a **supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**

b) In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

c) The decisions or the test are performed on the basis of features of the given dataset.

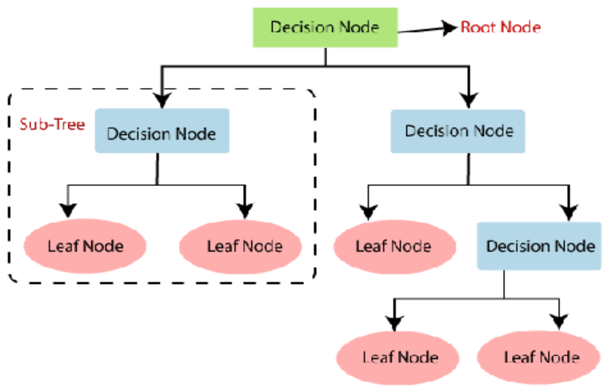
d) ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***

e) It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

f) In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**

g) A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

h) Below diagram explains the general structure of a decision tree:



4. RESULTS

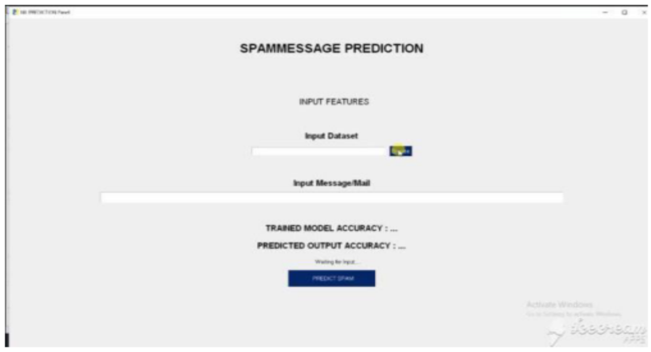


Fig 5.1: Browse the Input

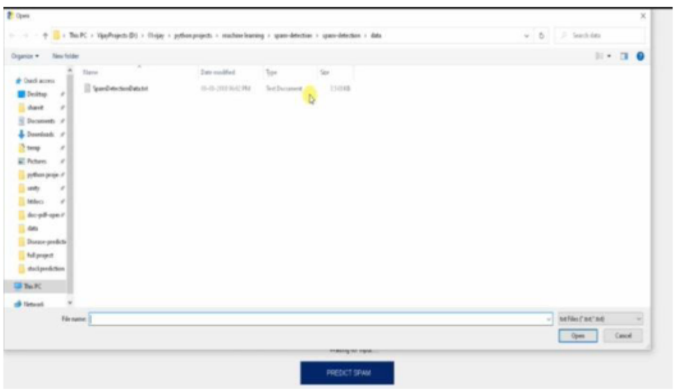


Fig 5.2: Browsing input from datasets

INTERNATIONAL JOURNAL OF RESEARCH AND DEVELOPMENT IN ENGINEERING SCIENCES
e-ISSN: 2582 - 4201 Volume: 2 | Issue: 6 | Year: 2020

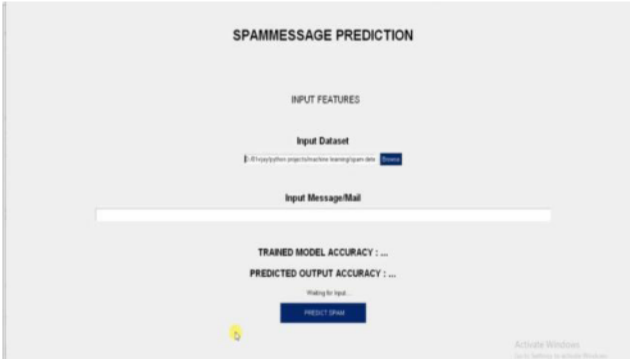


Fig 5.3: Input Data loaded

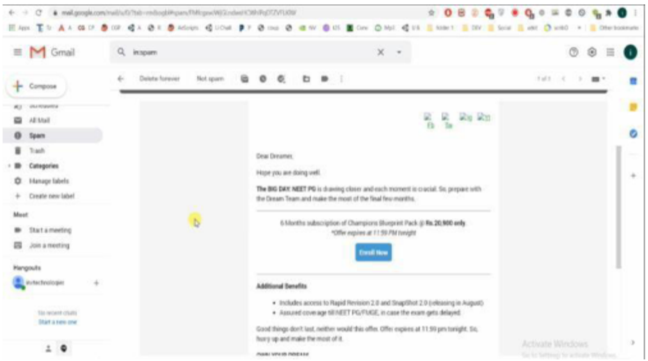


Fig 5.4: Input Message

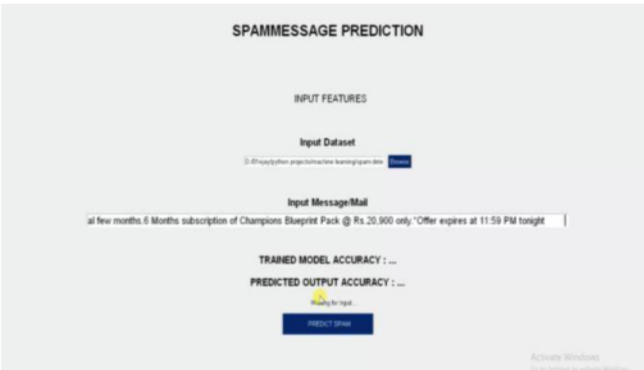


Fig 5.4: Data and Message are Loaded

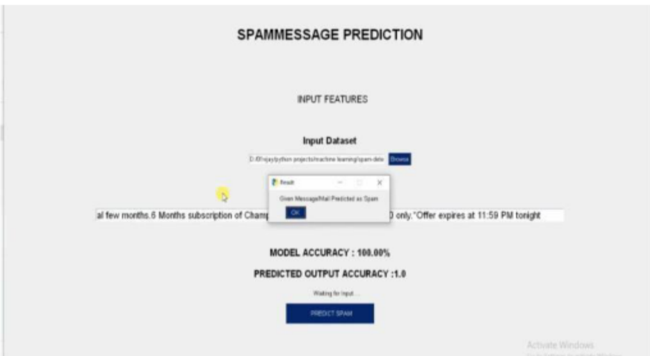


Fig 5.5: Output

5. CONCLUSION

To conclude, this paper has investigated the mathematical process of Naive Bayes Classifier in email servers, and how this algorithm applies the extended version of Bayes Theorem to compute the probability of each email being spam and further classify the emails into "Spam" section and "Inbox" section. Furthermore, this paper has explored the critical role threshold value play in this algorithm, and how does the variance of this value affect the final classification results.

REFERENCES

[1]. P.Sudhakar,G.Poonkuzhali, .Thaigarajan,K.Sarukesi, International Journal of Compuers, Issue 3, Volume 5, 2011, P. 332-35.

[2]. Almeida T, Yamakami A, Almeida J (2009) Evaluation of approaches for dimensionality reduction applied with Naive Bayes anti-spam filters. In: Proceedings of the 8th IEEE international conference on machine learning and applications, Miami, FL, USA, pp 517–522

[3]. Cormack G (2008) Email spam filtering: a systematic review. Found Trends InfRetr 1(4):335–455

[4]. Machine Learning Techniques in Spam Filtering Konstantin Tretyakov, kt@ut.ee, Institute of Computer Science, University of Tartu, Data Mining Problem-oriented Seminar, MTAT.03.177, May 2004, pp. 60-79.

[5]. A Study on Email Spam Filtering Techniques, Christina V et. all. *International Journal of Computer Applications (0975 – 8887) Volume 12– No.1, December 2010, pp.07-09*

[6]. Adaptive Spai Mail Filtering Using Genetic Algorithm,SanpakdeU et.all. Advanced Communication

[7]. Technology, 2006. ICACT 2006. The8th International Conference, 20-22Feb. 2006, Vol 1, 441 - 445

[8]. J.Quinlan. C 4. 5: Programs forMachine Learning. Morgan Kaufmann,1992.

[9]. V.Christina et al. Email Spam Filtering using Supervised Machine Learning Techniques. International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 09, 2010, 3126-3129 hmed Khorsi, "An Overview of Content-based Spam Filterin Techniques",*Informatica*, vol. 31, no. 3, October 2007, pp 269-277.

[10]. Weka. WEKA(DataMiningSoftware).Avail ableathttp:// www.cs.waikato.ac.nz/ml/weka/2006

[11]. C. Huang, "Research on SMS filtering technology on intelligent mobile phone," M.S. thesis, Huazhong University of Science and Technology, Wuhan, China, 2012.

[12]. J. Ma, Y. Zhang and Z. Wang, "A message topic model for multi-grain SMS spam filtering," *International Journal of Technology and Human Interaction (IJTHI),* vol. 12, no. 2, pp. 83-95, 2016.

[13]. S. J. Delany, M. Buckley and D. Greene, "Review: SMS spam filtering: Methods and data," *Expert Systems with Applications,* vol. 39, no. 10, pp. 9899-9908, 2012.

[14]. J. Fdez-Glez, D. Ruano-Ordas and J. R. Méndez, "A dynamic model for integrating simple web spam classification techniques," *Expert Systems with Applications,* vol. 42, no.21, pp. 7969-7978, 2015.

