International Journal ozf Gender, Science and Technology
ISSN: 2040-0748
UGC Care Group I Journal
Vol-12 Issue-01 2023

SMS SPAM DETECTION USING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

Dr.Vaka Murali Mohan[1]., S. Sadvika[2]., P. Deekshitha[3]., S. Rishik Reddy[4]., M. Sampath Reddy[5]

[1]Principal & Professor, [2,3,4,5]Students B.Tech -IT, (20S11A1230, 20S11A1213, 20S11A1229, 234).

Malla Reddy Institute of Technology and Science., Maisammaguda., Medchal., Ts, India,

[1]vakamuralimohan@gmail.com

,[2]sadvikasama@gmail.com,[3]deekshithapabboji@gmail.com,[4]reddyrishik2003@gmail.com,[5] sampathsubbu3@gmail.com

ABSTRACT

The number of people using mobile devices increasing day by day. SMS(short message service) is a text message service available in smartphones as well as basic phones. So, the traffic of SMS increased drastically. The spam messages also increased. The spammers try to send spam messages for their financial or business benefits like market growth, lottery ticket information, credit card information, etc. So, spam classification has special attention. In this paper, we applied various machine learning and deep learning techniques for SMS spam detection. we used a dataset from UCI and build a spam detection model. Our experimental results have shown that our Adaboost model outperforms previous models in spam detection with an accuracy of 98%. We used python for all implementations.

1.INTRODUCTION

The number of mobile phone (smartphone) users increases from 1 billion to 3.8 billion in five years . The top three countries using more mobiles are China, India, US. Short Message Service or SMS is a text messaging service available for the last several years. SMS service can be availed without internet also. So, SMS service is available in smartphones and basic mobiles also. Although smart phones bring several apps like WhatsApp for text messaging, this service can be availed with the help of the internet only. But SMS can be availed at any time. So, the traffic for SMS service increasing day by day. A spammer is a person/company which is responsible for unsolicited messages. For their organization benefits or personal benefits, spammers sending a vast number of messages to the users. These messages are called spam messages

2.LITERATURE SURVEY

Applying ML and DL techniques for spam detection is not a new era. Previously, various researchers applied ML techniques for classification SMS spam. Nilam Nur Amir Sjarif[3] et.al applied the TF-IDF technique in combination with a random forest classifier and achieved an accuracy of 97.5%.TF-IDF is a method used to quantify the words in a document by using two measures Term Frequency and Inverse Document Frequency.A.Lakshmanarao[4] et.al applied four machine learning classifiers Decision Trees, Naive Bayes, Logistic Regression, Random Forest for email spam filtering, and achieved an accuracy of 97% with random forest classifier. Pavas Navaney[5] et.al proposed various machine learning algorithms and achieved an accuracy of 97.4% with support vector machines. Luo GuangJun [6] et.al applied various shallow machine learning algorithms and achieved a good accuracy rate with logistic regression classifier. Tian Xia[7] et.al proposed the Hidden Markov Model for the detection of SMS spam. Their model used the information about the order of words thereby solving issues with low term frequency. They achieved an accuracy of 98% with their proposed HMM model. M. Nivaashini [8] et.al applied a deep neural network for SMS spam detection and achieved an accuracy of 98% They also compared DNN performance with NB, Random Forest, SVM, and KNN. Mehul Gupta[9] et.al compared various spam detection machine learning models with deep learning models and shown that deep learning models achieved a high accuracy rate in SMS spam detection. Gomatham Sai Sravya[10] et.al compared various machine learning algorithms for SMS spam detection and achieved the best accuracy with the Naive Bayes classification model. M.Rubin Julis[11] et.al applied various machine learning classifiers and achieved an accuracy of 97% with a support vector machine. K. Sree Ram Murthy [12] et.al proposed Recurrent Neural Networks for SMS spam

detection and achieved a good accuracy rate. S. Sheikh[13] proposed SMS spam detection using feature selection and the Neural Network model and achieved a good accuracy rate. Adem Tekerek[14] et.al applied various machine learning classification models for SMS spam detection and achieved an accuracy of 97% with a support vector machine classifier.

3.DATASET

The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged acording being ham (legitimate) or spam. A collection of 425 SMS spam messages was manually extracted from the Grumbletext Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages.

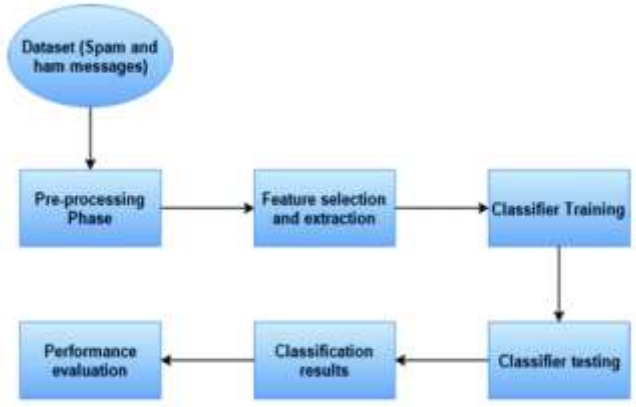
4.METHODOLOGY

Existing System

Although there are various SMS spam filtering techniques available, still there is a need to handle this problem with advanced techniques. Mobile users may get annoyed because of spam messages. Spam messages can be two types, SMS spam or email spam. The purpose of email spam or SMS spam is the same. Generally, these spam messages are spent by spammers for the promotion of their utilities or business. Sometimes, the users may also undergo financial loss due to these spam messages

Proposed System

Machine Learning is a technology, where machines learn from previous data and made a prediction on future data. Nowadays, machine learning and deep learning can be applied to solve most of the real-world problems in all sectors like health, security, market analysis, etc. There are various techniques available in machine learning like supervised learning, unsupervised, semi -supervised learning, etc. In supervised learning, the dataset is having output labels, whereas unsupervised learning deals with datasets with no labels. We used a dataset from UCI with labels, so we applied various supervised learning algorithms for SMS spam detection.



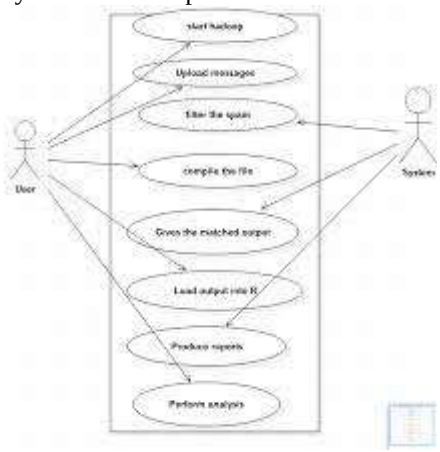
UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modelling language in the field of object-oriented software engineering. The standard is managed and was created by, the Object Management Group. The goal is for UML to become a common language for creating models of object-oriented computer software. In its current form, UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML. The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artefacts of software systems, as well as for business modelling and other non-software systems. The UML represents a collection of best engineering practices that have proven successful in the modelling of large and complex systems. The UML is a very important part of developing object- oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

USE CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality

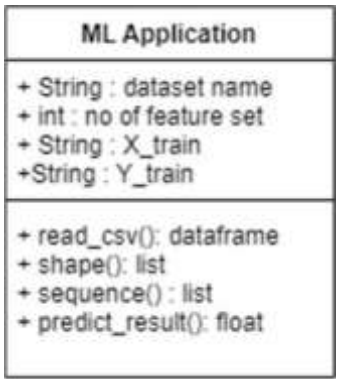
provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted



USE CASE DIAGRAM

CLASS DIAGRAM

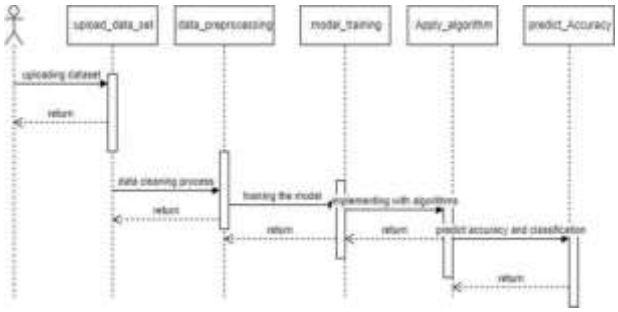
In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



CLASS DIAGRAM

SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



SEQUENCE DIAGRAM

SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead

to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub- assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail unacceptably. There are various types of tests.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expectedresults.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components

Functional testing

Functional tests provide systematic demonstrations

that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

System Testing

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration- oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Integration testing

Software integration testing is the incremental integration testing of two or more integratedsoftware components on a single platform to produce failures caused by interface defects.
The task of the integration test is to check that components or software applications, e.g., components in a software system or – one step up – software applications at the company level – interact without error.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participationby the end

user. It also ensures that the system meets the functional requirements.

5.ACKNOWLEDGEMENT

The team members of the research project want to sincerely thank our guide Principal & Professor Dr.Vaka murali Mohan and the Department of Information Technology, Malla Reddy Institute of Technology and Science, India for their encouragement and support for the completion of this work.

6.CONCLUSION AND FUTURE SCOPE

CONCLUSION

We proposed a deep learning model for SMS spam detection. We used a UCI dataset for our experiments. We applied three different word embedding techniques count vectorizer, TFIDF, Hashing Vectorizer. Later, we applied various classification algorithms. We achieved an accuracy of 98.5% with the LSTM model. Experimental results showed that our model outperforms previous models for spam detection

FUTURE SCOPE

Future scope of this project will involve adding more feature parameter. The more the parameters are taken into account more will be the accuracy. The algorithms can also be applied for analyzing the contents of public comments and thus determine 54 patterns/relationships between the customer and the company. The use of traditional algorithms and data mining techniques can also help predict the corporation performance structure as a whole. In the future, we plan to integrate neural network with some other techniques such as genetic algorithm or fuzzy logic. Genetic algorithm can be used to identify optimal network architecture and training parameters. Fuzzy logic provides the ability to account for some uncertainty produced by the neural network predictions. Their uses in conjunction with neural network could provide an improvement for SMS spam prediction.

7.REFERENCES

1. P. Mohan, D. Marin, S. Sultan, and A. Deen, “Medinet: personalizing the self-care process for patients with diabetes and cardiovascular disease using mobile telephony.” *Conference Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society*, vol. 2008, no.3, pp.755–758. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19162765
2. A. Tsanas, M. Little, P. McSharry, and L. Ramig, “Accurate telemonitoring of parkinson’s disease progression by noninvasive speech tests,” *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 4, pp. 884 – 893, 2010.
3. G. Clifford and D. Clifton, “Wireless technology in disease management and medicine,” *Annual Review of Medicine*, vol. 63, pp. 479–492, 2012.
4. L. Ponemon Institute, “Americans’ opinions on healthcare privacy, available: http://tinyurl.com/4atsdlj,” 2010.
5. A. V. Dhukaram, C. Baber, L. Elloumi, B.-J. van Beijnum, and P. D. Stefanis, “Enduser perception towards pervasive cardiac healthcare services: Benefits, acceptance, adoption, risks, security, privacy and trust,” in *PervasiveHealth*, 2011, pp. 478–484.
6. M. Delgado, “The evolution of health care it: Are current u.s. privacy policies ready for the clouds?” in *SERVICES*, 2011, pp. 371–378.
7. N. Singer, “When 2+ 2 equals a privacy question,” *New York Times*, 2009.

8. E. B. Fernandez, “Security in data intensive computing systems,” in *Handbook of Data Intensive Computing*, 2011, pp. 447–466.

9. A. Narayanan and V. Shmatikov, “Myths and fallacies of personally identifiable information,” *Communications of the ACM*, vol. 53, no. 6, pp. 24–26, 2010.

10. P. Baldi, R. Baronio, E. D. Cristofaro, P. Gasti, and G. Tsudik, “Countering gattaca: efficient and secure testing of fully-sequenced human genomes,” in *ACM Conference on Computer and Communications Security*, 2011, pp. 691–702.