VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering
VIT Chennai
Vandalur - Kelambakkam Road, Chennai - 600 127

Final Review Report

Programme: BTECH

Course: NATURAL LANGUAGE PROGRAM

Slot: E2

Faculty: DR. M. PREMALATHA

Component: J-COMPONENT

Title: EMAIL SPAM CLASSIFICATION USING SVM CLASSIFIER

Team Member(s): [Min 1, Max-3 in a team]

1. AAYUSHI (20BCE1791)
2. GARVIT JAKAR (20BCE1838)

Note:

1. Every explanation in Introduction, literature survey, Implementation, results and discussion, conclusion has to be cited in sequence from 1, 2, 3…..
2. A minimum of 15 journal papers has to be referred in the references and the content
3. Tables and figures have to be numbered and have to be specified with the title.

Abstract

Today, spamming email is one of the main difficulties encountered by everyone in the world of the Internet. In such a world, email is mostly shared by everyone to share the information and files because of their easy way of communication and for their low cost. But such emails are mostly affecting the professionals as well as individuals by the way of sending spam emails. Every day, the rate of spam emails and spam messages is increasing. Such spam emails are mostly sent by people to earn income or for any advertisement for their benefit. For those who receive the spam mail, the rising volume of it clogs up traffic and wastes their time. Sometimes, the users easily get trapped into financial fraud actions, by seeing the spam mails such as job alert mails and commercial mails and offer emails. The individual can experience some mental stress as a result. The system has presented a machine learning model to decrease all of these dangers. This model will identify spam and non-spam emails, and it will also optimise the data by deleting undesirable emails that contain advertisements, as well as some pointless and fraudulent emails.

Keywords : Natural Language Processing, Spam, Email, messages

1. Introduction

Email spam has grown significantly in recent years along with the rapid expansion of internet users. They are being used for fraud, phishing, and other unethical and criminal activities. sending harmful links through unsolicited email, which can damage our system and try to access your system. The spammers target those people who are unaware of these frauds and target them by easily creating phoney profiles and email accounts. In their spam emails, they pose as a real person. Therefore, it is necessary to identify spam emails that contain fraud. This project will do so by utilising machine learning techniques. This article will cover machine learning algorithms and apply all of them to our data.

Email spam detection using Support Vector Machines (SVM) is a popular and effective method for identifying unsolicited and unwanted emails. SVM is a supervised learning algorithm that uses labelled training data to classify new data points into one of two categories, in this case, spam or not spam. To build an email spam detection system using SVM, a dataset of labelled emails is first collected. The dataset is then pre-processed to extract relevant features, such as the presence or absence of certain keywords, the length of the email, and the frequency of certain characters. Next, the dataset is split into training and testing sets, and an SVM model is trained on the training set using the extracted features. The model is then evaluated on the testing set to determine its performance. Several techniques can be used to improve the performance of an SVM-based email spam detection system, such as feature selection, parameter tuning, and ensemble methods. Overall, SVM-based email spam detection is a robust and accurate approach for identifying and filtering unwanted emails from users' inboxes.

The main problems faced by all e-mail users are spam which contains unwanted information and data and some fake data which spoil people's lives and some messages which cause harmful effects. Today, 50 percent of people, both educated and uneducated, face job problems. In this case, these people receive emails of advertising messages that are completely fake. But when they see this post, those people are interested or have an idea to interact with the posts they are researching. This spam affects more people in similar cases. To reduce this risk and save people from the threat of spam, we recommend this spam removal system. We use two filter models to filter spam in this system.

2. Literature Review:

[1] The goal of this article is to identify and evaluate current state-of-the-art methodologies for SMS spam based on specific parameters, including the general acceptability of existing SMS applications and AI methods and techniques, approaches, and deployed environments. Mobile device technology has advanced steadily over time, providing consumers with an unrivalled communication experience that has effectively enhanced user performance . Information is now literally in everyone's hands because to the good effects that mobile devices have had on consumers [1] . Users haven't adjusted to the new ways of handling and understanding these devices, nevertheless. The Short Message Services, also referred to as SMS, are the most well-known and frequently used GSM service. Over 6 billion consumers worldwide have access to this quick-expanding service, which delivered 9.5 trillion SMS worldwide in 2009.[3]. As a result, SMS service has demonstrated its distinctiveness and validated its dependability as it continues to be one of the most popular communication tools among mobile users due to its simplicity, affordability, compatibility, and real-time services.

[2] Spam can be considered as any unsolicited, commercial, bulk electronic message (Graham, 2002) which may sometimes convey un-demanded adverts, viruses, malware or other annoying contents targeted at consumers, businesses or government organizations. Spam comes in different medium like email, SMS, Instant message (SPIM), Usenet newsgroup, social network, search engines, Internet Telephony, etc. (Abayomi-Alli, 2009). However, differentiating all these spam media is technically cumbersome for a review [7] reports SMS spam as not only offensive to users, but also incurs unnecessary cost on both MNOs and customers. The effects of SMS spam on users cut across different area of life, including financial, security/ privacy, education, career, health, social networking, etc.

[3] As business activity shifts from the old PC-based platform to the mobile phone platform, mobile security has recently become a top concern (Lau, 2017). Unlike SMS spam, which is constantly increasing, email spam challenges have shown tremendous progress and improvement over time [3]. Although the use of artificial

intelligence (AI) in automatic text classification is largely positive, there is still need for improvement when it comes to brief text classification, particularly for SMS and microblogging (Huang et al., 2018) Thus, the ongoing and obnoxious growth of brief text messages in mobile devices serves as the inspiration for this systematic literature review (SLR). Additionally, research papers on SMS spam that have already been published have not been able to support the many approaches researchers have taken to address the SMS spam problem. In order to give future researchers the opportunity to exploit underutilised or unused methodology with the goal of improving accuracy and overall system performance, this study aims to identify all possible measures (methods/techniques, approaches, architecture) adopted by researchers in SMS spam classification.

[4] Even though research is being done to combat SMS spam, there is still much that needs to be done to improve the classifier that is now in place for the better classification of obscured short messages. This article gives a comprehensive review for classifying SMS spam with the goal of finding and analysing the most cutting-edge approaches used to do so. Due to a sizable number of difficulties encountered in actual implementation, SMS spam solutions have scarcely been used on mobile devices or by mobile network operators (server). Although customers appear to have taken to the arrival of other messaging services on mobile devices, the importance of SMS in terms of dependability, cost-effectiveness, internet independence, etc. cannot be overstated.

[5] The growth of mobile phone users has led to a dramatic increase in SMS spam messages. Despite the fact that a variety of information channels are currently seen as "spotless" and reliable in many parts of the world, ongoing data clearly demonstrate that the amount of cell phone spam is dramatically increasing over time. It is a growing catastrophe, especially in the Middle East and Asia. Separating SMS spam is a similarly late task to solve this problem. It gains several concerns and practical fixes from SMS spam separation. In any case, it brings up its own unique problems. By including Indian messages in the entire available SMS dataset, this research aims to address the challenge of classifying flexible messages as Ham or Spam for the Indian Users. The research examines various machine learning classifiers on a sizable dataset of individual SMS texts.

[6] Mobile message is a way of communication among the people, and billions of mobile device users exchange numerous messages. However, such type of communication is insecure due to lack of proper message filtering mechanisms. One cause of such insecurity is spam, and it makes the mobile message communication insecure. Spam is considered to be one of the serious problems in e-mail and instance message services. Spam is a junk mail or message. Spam e-mails and messages are unwanted for receivers which are sent to the users without their prior permission. It contains different forms such as adult content, selling item or services, and so on [1]. The spam increased in these days due more mobile devices deployed in environment for e-mail and message communication. Currently, 85% of mails and messages received by mobile users are spam [2]. The cost of mails and messages are very low for senders but high for receipts of these messages. The cost paid some time by service providers and the cost of spam can be measured in the loss of human time and loss of important messages or mails [3]. Due to these spam mails and messages, the

values able emails and messages are affected because each user have limited Internet services, short time, and memory [4].

[7] The way we live has changed or been revolutionised by mobile or smart phones. Short message service (SMS) is growing in popularity these days. The popularity of the mobile messaging platform has made it a very alluring target for spammers to attack. We will design a system that is more authenticated for SMS in order to impose an additional layer of security in the pervasive environment. From the perspective of the user's safety, this system will affect how easy it is for users to use. The financial sector and other related organisations now view SMS as a crucial component of customer communication, which inadvertently creates an opening for spammers and puts customers' safety measures in jeopardy. The SMS formation, which requires two nodes to swap over digitally signed SMS messages, benefits from the use of digital encryption methodologies. Public key cryptography is used to secure these two nodes, and the ECDSA signature scheme is used for authentication. These two nodes are identified as the sender and the receiver, and whenever a sender sends an SMS to a recipient, the unencrypted text is sent, which increases the risk of data loss. In this article, we suggest the Gaussian Naive Bayes Classification (GNBC) method for SMS spam filtering that addresses issues with message topic models (MTM).[4] Spammers target SMS and email messages as well as other domains with spam in order to steal data, money, and other things. There are a number of models for detecting spam in SMS and emails, however the majority of them rely on supervised learning. However, a thorough study for spam detection that takes into account numerous domains at once is lacking. In this study, the effectiveness of feature selection strategies is experimentally evaluated in the context of SMS and email spam detection. Performance of spam detection models is assessed using standard parameters as well as factors like ROC and Train/Test time. The experiment's findings demonstrate that the feature selection method selected has a significant impact on the spam detection model's effectiveness, as evidenced by the results of many assessment measures, some of which have not been applied to the two domains before.

[8] In [6], the e-mail classification method was proposed for the detection of spam. In the system, four predictive machine learning classifiers were used with various data partitions for training and testing of the models. Additionally different hyper parameters values were used in the models. The system obtained good results. Bhat [7] designed ensemble methods based on techniques such as bagging, boosting, and stacking for classification of spam and ham. The data set used in the study was collected from Facebook. The experimental results demonstrated that the bagging ensemble learning approach, using J48 (decision tree) base classifier, performs well than its individual model, and the method achieved high performance in terms of detection accuracy. In [4], a method is proposed for ham and spam detection and principle components analysis and support vector machine were used in the designing of the system. Additionally, the performance evaluation and cross validation methods were used in the system. The proposed technique achieved high performance, and the method effectively detected the spam.

[9] Spammers target SMS and email messages as well as other domains with spam in order to steal data, money, and other things. There are a number of models for detecting spam in SMS and emails, however the majority of them rely on supervised

learning. However, a thorough study for spam detection that takes into account numerous domains at once is lacking. In this study, the effectiveness of feature selection strategies is experimentally evaluated in the context of SMS and email spam detection. Performance of spam detection models is assessed using standard parameters as well as factors like ROC and Train/Test time. The experiment's findings demonstrate that the feature selection method selected has a significant impact on the spam detection model's effectiveness, as evidenced by the results of many assessment measures, some of which have not been applied to the two domains before.

[10] "A Comparative Study of Machine Learning Techniques for Email Spam Classification" by Gokila et al. This paper compares the performance of different machine learning algorithms, including Naive Bayes, Support Vector Machines, and Random Forests, for email spam classification.

[11] The experimental results ,the classification performance of LR is high as compared with the decision and k-nearest neighbour in terms of accuracy. Similarly, the computation time of LR is low as compared with k-NN and DT.. From these experiential results analysis, we concluded that the LR effectively classifies the ham and spam because the achieved accuracy is high. The 100% specificity of the LR model correctly detected the ham messages. Similarly, 86% sensitivity shows that LR spam message capability is good. Thus, the experimental results suggest that LR is a the best classifier for the classifications of ham and spam successfully. This paper moreover focuses on the importance and application of using Machine learning Algorithms to classify spam emails and SMS.

[12] "A Survey on Email Spam Filtering Techniques" by Jyoti Agrawal et al. This paper provides an extensive survey of email spam filtering techniques, including rule-based, content-based, and machine learning-based methods. The authors also discuss the challenges and future directions in this field.

[13] "A Review of Email Spam Filtering Techniques" by Anand Deshpande and Shailendra Singh Thakur. This paper provides an overview of the various techniques used for email spam filtering, including rule-based, content-based, and machine learning-based methods.

[14] "Spam Filtering with Naive Bayes – Which Naive Bayes?" by Paul Graham. This paper proposes an improved Naive Bayes algorithm for email spam filtering, which takes into account the frequency of words in spam and non-spam messages.

[15] "An Ensemble Approach to Email Spam Filtering using Machine Learning Algorithms" by Muhammad Usman et al. This paper proposes an ensemble approach to email spam filtering, which combines the outputs of multiple machine learning algorithms to improve classification accuracy.

3. **Data Set Description**:

Dataset used :- Messages.csv

https://github.com/priyalagarwal27/E-mail-spam-detection/blob/main/messages.csv

It contains 3 columns and 2964 rows with column values named subject, message and label. Subject refers to the topic of the email that is being talked about in the corresponding message section and label is an important column as it contains binary values i.e. 0 and 1 . 0 stands for NOT SPAM and 1 stands for SPAM.
The dataset undergoes certain pre-processing steps to turn it into a clean dataset.
The various pre-processing steps are :-

1. Read dataset and make it in proper format
2. Encode labels : 0 and 1
3. Convert all cases to lower
4. Remove punctuations
5. Remove stop words

Tokenisation – Tokenization is the process of breaking down a piece of text into individual words or groups of words, called tokens. In natural language processing (NLP), tokenization is a crucial step for text processing because it allows the machine to understand and process the meaning of text more effectively.

Lemmatization – Lemmatization is the process of reducing a word to its base or root form, known as a lemma, which represents the canonical form of the word. In natural language processing (NLP), lemmatization is an important technique used to standardize text and reduce variation for downstream analysis.

4. **Implementation:**

Email spam detection system is used to detect email spam using Machine Learning technique called Natural Language Processing and Python, where we have a dataset contain a lot of emails by extract important words and then use naive classifier we can detect if this email is spam or not.

Project Pipeline

For any machine learning project, it consists of three main phases as following: -

1. Scoping: List the problem description and project goals.

The project problem is that we have a dataset containing a set of emails and we will use machine learning and NLP techniques in order to determine if this email is spam or not.

2. The Data: Load, analyse and prepare the dataset for training.

In this phase we will analyse and prepare the dataset before training by applying various steps as following: -

- Data Loading
- Data Visualization
- Data Cleaning
- Data Splitting

Data Visualisation:

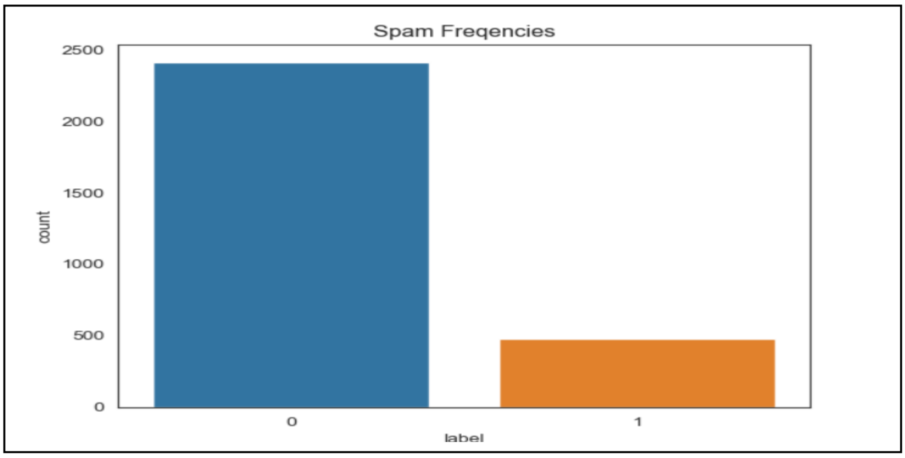


Fig 1. Checking the spam frequencies for both the levels 0 and 1

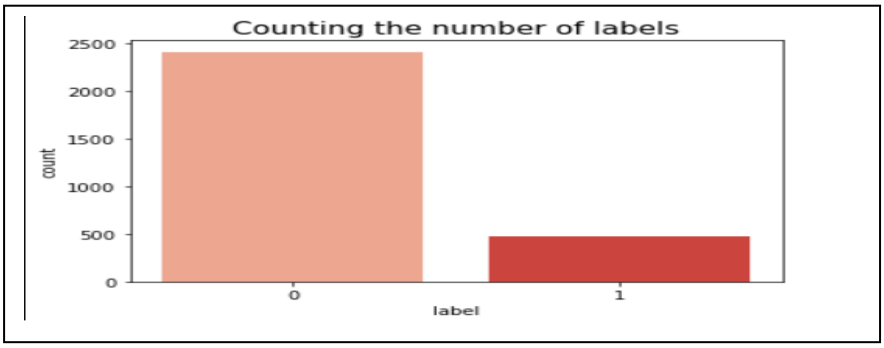


Fig 2. Counting the number of labels for every text

3. The Model: Create and train the model on the dataset

Models used: -

1. **Naïve Bayes :-** Naive Bayes is a classification algorithm based on Bayes' theorem, which states that the probability of a hypothesis (class) is proportional to the probability of the evidence (features) given that hypothesis, multiplied by the prior probability of that hypothesis. In the context of classification, the naive Bayes algorithm assumes that the features are conditionally independent given the

class, which means that the presence or absence of one feature does not affect the probability of another feature occurring, given the class. This simplifies the computation of the probabilities, and allows the algorithm to work well even with a small number of training samples.

2. **Support Vector Machine (SVM) classifier :-** Support Vector Machine (SVM) is a popular supervised learning algorithm used for classification and regression analysis. In classification, SVM is a binary classifier that tries to find a hyperplane in a high-dimensional space that maximally separates the different classes. The hyperplane is chosen such that it maximizes the margin between the classes, which is the distance between the hyperplane and the nearest data points from each class. The SVM algorithm tries to find the hyperplane that maximizes the margin between the classes, while also minimizing the classification error. The optimization problem is formulated as a quadratic programming problem, which can be solved efficiently using standard optimization techniques. SVM is a powerful algorithm that can handle complex, high-dimensional data and is widely used in various applications such as text classification, image classification, and bioinformatics. SVM can also be extended to handle multi-class classification problems and regression analysis.

Model Evaluation: -

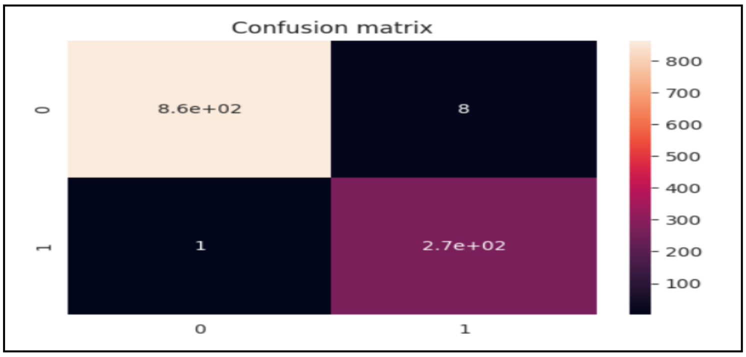
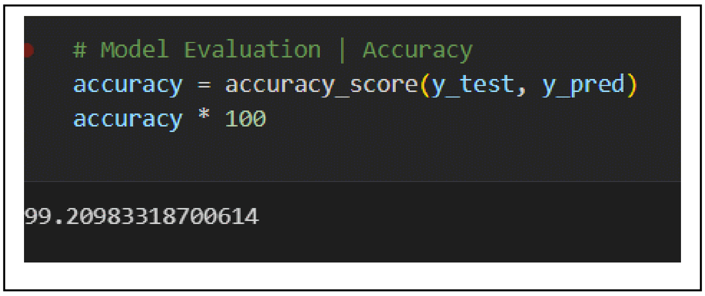


Fig 3. The model accuracy is evaluated using a confusion matrix

4. **Results and Discussion:** [Explain in detail about the results obtained from the models applied. Specify all the metrics, compare the results of the models and justify the same]
In naïve bayes the accuracy score was in between 83 to 88 % & in SVM the accuracy score is around 98%.



```
# Model Evaluation | Accuracy
accuracy = accuracy_score(y_test, y_pred)
accuracy * 100
```

```
99.20983318700614
```

Fig 4. Accuracy of proposed model is 99.2098%

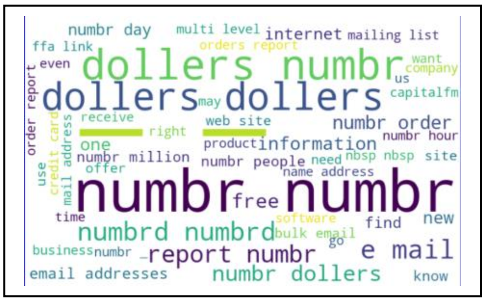


Fig 5. Word cloud model of word count in the dataset

Classification Report of SVM combined with NLP:-

	Precision	Recall	F1 Score	support
Accuracy	0	0.99	1.00	585
Macro average	0.99	0.94	0.97	139
Weighted average	0.99	0.97	0.98	724

5. **Conclusion:**

E-mail spam filtering is an important issue in the network security. Naïve Bayes classifier that used has a very important role in this process and its quality of performance is also based on datasets that is being used. Naïve Bayes classifier also can get high precision that give high percentage of spam message manage to block if the dataset collect from single e-mail accounts. In our project, Naïve Bayes gave the accuracy of around 86%. This model either lack in performance of level of accuracy.

New mechanism using SVM is proposed to enhance the precision of the spam detection. SVM classifier get the highest precision that give highest percentage of spam message manage to block. Support Vector Machine (SVM) gave accuracy of 98%. The results are comparatively better the naïve bayes classifier.

6. Future Work:

In order to minimise network traffic and data storage, this model could be changed to operate on the sender side rather than the receiver side. Additionally, email IDs could be ranked; by doing so, the aforementioned issues could also be resolved. The alternative techniques include storing only the header, attachments, and links for analysis in place of the entire message. A method to encrypt the private messages selected by the sender could be used or the aforementioned point could be used to protect an individual's privacy. The model's dataset could be updated to reflect current patterns for greater accuracy.

7. References

1. Abu-Nimeh S, Nappa D, Wang X, Nair S (2008) Bayesian additive regression trees-based spam detection for enhanced email privacy. In: 2008 third international conference on availability, reliability and security. IEEE, pp. 1044–1051. doi:10.1109/ARES.2008.136
2. Adewumi AAAA, Owolabi TO, Alade IOIO, Olatunji SO (2016) Estimation of physical, mechanical and hydrological properties of permeable concrete using computational intelligence approach. Appl Soft Comput 42:342–350. doi:10.1016/j.asoc.2016.02.009
3. Akande KOKO, Owolabi TO, Olatunji SO (2015) Investigating the effect of correlation-based feature selection on the performance of support vector machines in reservoir characterization. J Nat Gas Sci Eng 22:515–522. doi:10.1016/j.jngse.2015.01.007
4. Akande KO, Olatunji SO, Owolabi TO, AbdulRaheem A (2015a) Comparative analysis of feature selection-based machine learning techniques in reservoir characterization. CPAPER, Society of Petroleum Engineers. doi:10.2118/178006-MS
5. Akande KO, Olatunji SO, Owolabi TO, AbdulRaheem A (2015b) Feature selection-based ANN for improved characterization of carbonate reservoir. CPAPER, Society of Petroleum Engineers. doi:10.2118/178029-MS
6. Akande KO, Owolabi TO, Twaha S, Olatunji SO (2014) Performance comparison of SVM and ANN in predicting compressive strength of concrete. IOSR J Comput Eng 16(5):88–94
7. Ariaeinejad R, Sadeghian A (2011) Spam detection system: a new approach based on interval type-2 fuzzy sets. In: 2011 24th Canadian conference on electrical and computer engineering(CCECE). IEEE, pp. 000379–000384. doi:10.1109/CCECE.2011.6030477
8. Cortes C, Vapnik V (1995) Support vector networks. Mach Learn 20:273–297
9. Fernandez R, Picard RW (2002) Dialog act classification from prosodic features using support vector machines. In: Speech Prosody. Conference paper, Aix-en Provence, France, Dialog Act
10. Gupta SM (2007) Support vector machines based modelling of concrete strength. World Acad Sci Eng Technol 36:305–311
11. Ibitoye M, Hamzaid N, Abdul Wahab A, Hasnan N, Olatunji S, Davis G (2016) Estimation of electrically-evoked knee torque from mechanomyography using support vector regression. Sensors 16(7):1115. doi:10.3390/s16071115
12. Idris I, Selamat A (2014) Improved email spam detection model with negative selection algorithm and particle swarm optimization. Appl Soft Comput 22:11–27. doi:10.1016/j.asoc.2014.05.002
13. Hopkins M, Reeber E, Forman G, Suermondt J (1999) SpamBase dataset. Hewlett-Packard Labs; 1501 Page Mill Rd.; Palo Alto; CA 94304. https://archive.ics.uci.edu/ml/datasets/Spambase

14. Ibitoye M, Hamzaid N, Abdul Wahab A, Hasnan N, Olatunji S, Davis G (2016) Estimation of electrically-evoked knee torque from mechanomyography using support vector regression. Sensors 16(7):1115. doi:10.3390/s16071115
15. A Sharaff and Srinivasarao U (2020), "Towards classification of email through selection of informative features," First International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, pp. 316-320, DOI: 10.1109/ICPC2T48082.2020.9071488.