International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Email Spams via Text Mining using Machine Learning Techniques

Tarika Verma, Nasib Singh Gill

Abstract: ***A lot of data is generated on daily basis which may potentially be useful. This data is generally unstructured and ambiguous to draw a meaning from it. High quality of information can be extracted from this potentially useful data typically through devising of patterns and trends in it. This is done using Text Mining which includes the initial parsing of the unstructured data, processing it and then leading to some meaningful and fascinating information hidden in it. This paper presents the machine learning techniques for text mining that are useful for spam detection in emails.***

Keywords: Text Mining, Machine Learning Techniques, Spam Mail Detection, ML Classifiers

I. INTRODUCTION

Text Mining is the field that seeks to elicit meaningful information from natural language text. It scrutinizes text to elicit information that is useful for an eccentric purpose. As compared with data type stored in DBs, text is irregular, ambiguous, and tough to process [1].

Nowadays every other person is involved in some activity that may be offline or online due to which some corresponding unstructured data may be generated. For example, web-based applications feed the web servers for the behavior of its users, person involved in some shopping might be buying certain specific type of products at a time, specific type of articles or news may be read by a certain group of people, certain type of emails or texts may be classified as spam by its users and so on. Thus, this particular behavior of a person may potentially be used for devising information from the collection of such data. This inference is drawn based on some patterns found in the available sample data and which can further be used in making predictions or taking decisions in future [2]. This is done in Text Mining using the Machine Learning Techniques.

The goal is, indeed to convert unstructured text into structured data format for analysis purpose, via the utilizing "natural language processing" (NLP).

Text mining is a flourishing field that elicits meaningful information from text of a human like language or natural language text. It can be distinguished as the process of scrutinizing text to evoke potential information that is helpful for an eccentric purpose. Machine learning (ML) is the scientific and the statistical study in which computers are used to draw inference regarding a task without being given the explicit instructions by the programmer.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Ms. Tarika Verma*, Assistant Professor, Computer Science and Engineering, AIJHM College, Rohtak, India.
Dr. Nasib Singh Gill, Professor, Dept. of CS & Applications, M. D. University, Rohtak, India.

This inference is drawn based on some patterns found in the available sample data and which can further be used in making predictions or taking decisions in future [2].

Fig. 1: Text Mining Process [3]

II. EMAIL SPAM DETECTION

The motivation of Email was the multiple users from remote locations, joining a central system to store and share data and documents from distal terminals. The terrific growth of email is due to its negligible cost, high efficiency, and concordance with many information types. Email is now omnipresent communication approach [4].

Since its commencement, emailing has speeded-up global operations to the heights of economic growth. It is so omnipresent in our regular lives that global figure, of daily sent emails, has reached 205 billion [6]. However, email protocols like "SMTP" and "POP" being easy to use and handy to everyone, increases their risk of being misused. Usually, plenty of mails are irrelevant, of no use, and un-demanded, which are normally auto-generated daily.

Five major email mining tasks are as shown in Fig 2. [5]

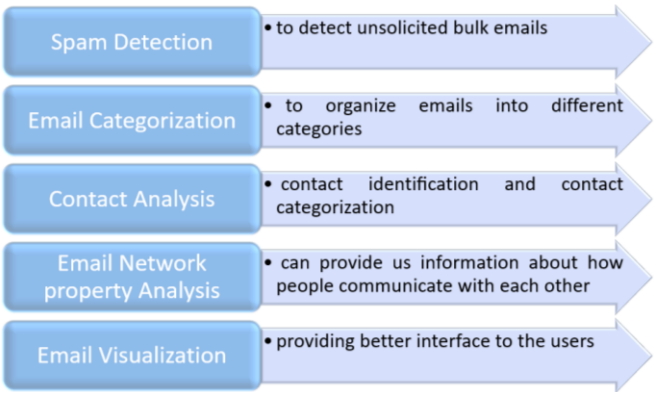

Fig 2: Email Mining [5]

Such emails are used for publicizing, crypto-worms, stealing user credentials, fake purchase bills, increasing web traffic to spiteful websites, loading malicious softwares, crimeware etc.

Due to this a lot of time is spent wastefully in handling these spams which cause 20 million dollars annual loss approximately [6].

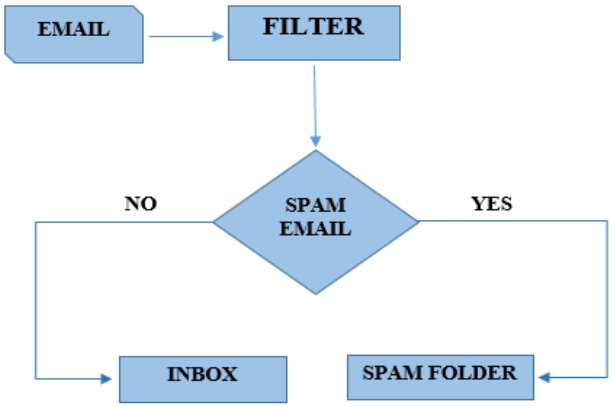

Fig 3: Generalized Flow Chart for Spam Mail Detection [7]

The spam arrival also incurs memory wastage on servers that adds up to extra cost to either the user or to the company even being of no use completely. This leads to purchasing of additional extra storage after a period of time. The storage size gets compounded exponentially with same email client being used by many users. [6]

Due to spams, important mails may be overlooked or may be deleted accidentally. As an important and common means of communication at each level of an enterprise, everyone depends upon email and the presence of spam influences an organisation on all levels. [6]

A. Effects of Spam

Besides being annoying, spam is also risky to its users. Spam mail are undesired and not requested by the users. These are normally sent to many people with malicious intent. Source or the sender can be anonymous or masked by some false mailing address. Additionally, no provision is available to unsubscribe such fake mails. [8]

Spam's negative consequences [7]:

• *Direct Impacts*

Spam is now avenue to trade in cheap goods, to plant malwares and viruses, to con people and so on. These directly impacts on the victims. The victim's computer can be used for cyber-attacks or other malicious activities. Additionally, the victim's imposition can be used in false or criminal activities.

• *Network's Resource Exhaustion*

Spam leads to email traffic. This traffic exhausts network storage and bandwidth. This leads to improper delivery or packets loss in transmission.

• *Human Resource Exhaustion*

Along with network bandwidth exhaustion, spams require a lot of human efforts in context of time wastage that happens due to the time spent in distinguishing normal mails from spams. The effects are increased even more when ham mails look like spam.

B. Machine Learning Classifiers

Spam mails are extensively spreading out daily and these cause a exigent loss. To prevent these spams many ways are present including ML [8].

Following are a few ML classifiers: [5]

- **Support Vector Machine**

These are binary classifiers which in emails context segregates emails into two classes ("spam" and "non spam") using hyperplane. That hyperplane is aimed for which can enlarge the gap between the two classes. [9]

- **Naïve Bayes**

These classifiers are omni present in spam detection area. These presume that the features values to be statistically independent. Email terms are extracted as features and the precision can be improved by attaching more features, like considering mail attachments and sender's domain in address. [10]

- **Decision Tree**

It is a "divide-and-conquer" approach, which creates a learning problem from a given self-reliant instances set. Here, the tree's root node depicts a condition or problem statement which has one or more solutions. Each solution further rises a problem set that resolves to the final. [11]

- **Linear Regression**

In it, some continuous quantity or variable is predicted. This is usually done by visualizing the relation between the dependent and independent variables. For eg. Predicted price vs Actual Price of some product over a period of time. In this we have to predict the dependent variable (Y) value on the basis of the dependent variable (X). It is generally used while predicting a continuous quantity. This dependent variable is always continuous in Regression Model. [12] The independent variable can be discrete or continuous. It can be represented by:

$$Y = b_0 + b_1 X + error$$

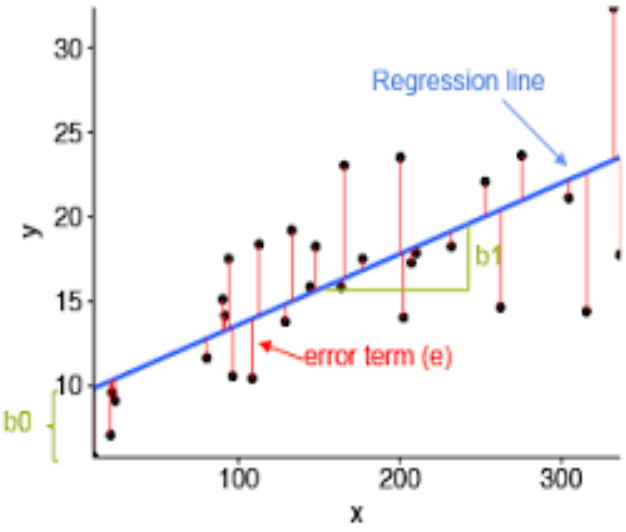

Fig. 4: Linear Regression [13]

- **Association Rule-based**

It is based on of "IF-THEN" rules instead of feature vectors. [14] For example:
IF "word FREE appears in subject" OR "word !!!! appears in subject"
THEN "the email is spam".
Here we define two terms:

- False alarm rate $= \frac{\text{Missclassified spams}}{\text{Total spams}}$
- Miss rate $= \frac{\text{Missclassified non spams}}{\text{Total non spams}}$

III. LITERATURE SURVEY

Using the machine learning algorithms under a supervised environment will train the system to classify emails as spam or ham. Also, as the training data is provided by the user, so the system is customized according to the user requirements for the classification of emails. This increases the security and productivity of the user.
These are some scholarly works related to this domain:
Aakash Atul Alurkar et al. (2017) proposes the emails classification using a ML approach which automatically recognizes the necessary features using various parameters more accurately. Its aim is to group main emails plus it blocks the spams. [6]
N Shajideen and Bindu V (2018) in [15] have presented ML classifiers SVM (Support Vector Machine), NB (Naïve Bayes) and J48 and also evaluated various parameters. This found SVM's False Positive Rate and accuracy are best.
Linda Huang, Julia Jia et al. (2018) enhanced Naive Bayes Spam Filter's accuracy and also implemented spam encryptions. They also analyzed challenges for individuals and companies by spam mails. [16]
S. Bhalero and M. Dalal (2017) have redesigned the SOAP [10] (Social network Aided Personalized and effective spam filter) method based on RBF (Radial Basis Function) neural classifier to represent the better spam filtering technique ISOAP (Improved SOAP) which out performs SOAP. [17]
Wanqing You et.al (2015) presented content based anti-spam filter using Enron Spam Dataset and "Naïve Thomas Bayes" technique. [18]
N Shajideen and Bindu V (2018) discussed a new ontology-based spam filtering method which prioritize personal interests and spam emails are classified on the basis of user profile preferences. Unlike conventional techniques, where the users don't have mail access control. [19]
In [20], **G. Caruana and M. Li (2012)** presented various computing application like "peer-to-peer computing", "grid computing", "semantic web" and "social networking" for spam filtering.
In [21], **R. Shams et al. (2013)** utilizes text features by its frequency and HTML tags for spam detection. They introduced "language centric features" such as grammar and errors in spelling, noticing alpha-numerics and verbs and "inverse sentence frequency". They used – "Random Forest, BAGGING, ADABOOSTM1, Support Vector Machine and Naive Bayes".
In [22], **N. O. F. Elssied et al. (2014)** considered "one-way ANOVA", "F-test" as a feature selection and "SVM based on poly kernel" as spam classifier.
K. Kowsari, D. E. Brown, et al. (2017) [23] did the hierarchical classification using "Hierarchical Deep Learning for Text classification" (HDLTex) which uses deep learning architectures and thus specialized understanding is provided at each document hierarchy level.
S. Saha et al. (2019) classified using spam mail using: "Naïve Bayes, SMO, J48, and random forest" in the 4601 instances data set. Classifiers are analysed and compared depending on their performance leading to "random forest technique" with max-accuracy, max-weighted precision, max-weighted recall, and max-weighted F-measure of 95.50. On the basis of execution time metrics Naive Bayes performs best. [24]
A. Barushka and Petr Hajek (2019) used "word embedding methods" to achieve better results in "review-spam detection". As per results the proposed "DNN and content-based approach" has best accuracy. [25]
A. Barushka and Petr Hajek (2018) in [26], proposed a spam filter that outperformed various approaches including "Minimum description length", "Factorial design analysis using SVM and NB", "Incremental learning with C4.5 decision tree", "Voting", "Random Forest" and "Convolutional neural network". This out performance was observed on all under-observation-datasets. This leads us to the fact that deep NNs is a promising spam filter technique. The results additionally proposed that increasing so many units and hidden layers would introduce training data noise and finally causing poor generalization in performance.
A. Barushka and Petr Hajek (2018) in [27] depicted "ensemble learning algorithms with DNN" as the base learner is more accurate than "state-of-the-art spam filtering methods". From results we know that "bagging algorithm trained with DNNs" achieved high accuracy and best results on both classes. This is attributed to the "bagging" capacity in reducing the over-fitting risk.
Maryam Shuaib et al. (2019) in [28] proposed the use of a "meta-heuristic optimization algorithm", the "whale optimization algorithm" (WOA), for the features selection in the email corpus and "rotation forest algorithm" for classifying email spams. Complete datasets were used, and the "rotation forest algorithm" evaluation was done afore and after feature selection.
Nida Mirza et al. (2017) tried to find "data mining techniques" based best spam-classifier. Bayesian Naïve Classifier is used and word extraction is done using the word count algorithm. According to results, "Naïve Bayesian Classifier" produces a better solution than "Support Vector Machine" [29].
Maria Habib et al. (2018) proposed spam detection based on "Genetic Programming" (GP) combined with "Synthetic Minority Over-sampling Technique" (SMOTE). It is applied and hence tested on two benchmark email corpora. Then it is tested on four classifiers using four measures: "accuracy, recall, precision and G-mean". As per the results "GP combined with SMOTE" can effectively do spam classification outperforming common classification methods. [30]
Sunday Olusanya Olatunji (2017) in [31] proposed email spam detector based on "SVM classifier". It is trained and then tested employing popular and standard database. This spam detector comes up with 3.11% improvement over the "negative selection algorithm" (NSA) with "particle swarm optimization" (PSO) i.e. NSA–PSO hybrid scheme.

V. Gupta et al. (2018) introduced ensemble learning technique in [32] to detect textual-spam. In this method, "voting classifier" is used and comparison with different "supervised and unsupervised classifier" is done. Final result shows that the max-accuracy is acquired when "decision tree, Gaussian Naive Bayes classifier, and Bernoulli Naive Bayes classifier" are used in "voting classifier".

IV. CONCLUSION

Email has become indispensable in our lives. Spam filtering is email prioritization kind that concentrates on classifying emails as spam and ham. This is important at both organizational level and individual level. For the organization it reduces burden on the server and increases the trustworthiness for the organization among its users. For an individual a secure email client is always desirable that ensures greater security. This paper presents several techniques that may be used for spam classification as well as provide the related work of various contemporary scholars on the spam classification.

V. RESULT

This paper presents various contemporary work in "spam filtering in emails" along with the prevalent "ML techniques". Email is now omnipresent communication approach which connects distal terminals to allow them sharing document and other data. But its protocols being easy and handy makes it prone to be misused. Also, the total email data is generated in a huge quantity on daily basis. Further there is the need to convert unstructured text into structured data format for analysis purpose. Afterwards, spam filtering is used for email prioritization to classify emails using "ML techniques" in order to save resources and to avoid time wastage.

REFERENCES

1. L. Kumar, P.K. Bhatia (2013). "Text mining: concepts, process and applications", Journal of Global Research in Computer Science, Vol 4, No. 3, March 2013. http://www.rroij.com/open-access/text-mining-concepts-process-and-applications-36-39.pdf
2. Online, Text mining, https://en.wikipedia.org/wiki/Text_mining, Accessed on: 26 Nov 2019
3. Online, https://data-flair.training/blogs/text-mining/, Accessed on: 10 Jan 2020
4. Email Statistics Report, 2015-2019 Executive Summary http://www.radicati.com/wp/wp-content/uploads/2015/02/Email-Statistics-Report-2015-2019-Executive-Summary.pdf
5. Tang, G., Pei, J. & Luk, WS. "Email mining: tasks, common techniques, and tools", Knowl Inf Syst (2014) 41: pp 1-31, https://doi.org/10.1007/s10115-013-0658-2
6. A. A. Alurkar et al., "A proposed data science approach for email spam classification using machine learning techniques," 2017 Internet of Things Business Models, Users, and Networks, Copenhagen, 2017, pp. 1-5. doi: 10.1109/CTTE.2017.8260935
7. Cormack, Gordon. (2006). "Email Spam Filtering: A Systematic Review". Foundations and Trends in Information Retrieval. 1. 335-455. Doi: 10.1561/1500000006.
8. Shajideen, Nasreen M and V Bindu. "Spam Filtering: A Comparison Between Different Machine Learning Classifiers." 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) (2018): 1919-1922.
9. Pham, B. T., Jaafari, A., Prakash, I., & Bui, D. T. (2018). "A novel hybrid intelligent model of support vector machines and the MultiBoost ensemble for landslide susceptibility modeling". Bulletin of Engineering Geology and the Environment. doi:10.1007/s10064-018-1281-y
10. Tarika Verma et al. (2017). "Enhancing K-Means and Naive Bayes for Data Mining". International Journal of Engineering and Technology. 348-350. Doi 10.21817/ijet/2017/v9i3/170903S053.
11. Tarika Verma, Chhavi Rana. (2017). "Data Mining Techniques for the Knowledge Discovery". International Journal of Engineering and Technology. 9. 351-354. Doi: 10.21817/ijet/2017/v9i3/170903S054.
12. [Online], Jason Brownlee (2016), "Linear Regression for Machine Learning", https://machinelearningmastery.com/linear-regression-for-machine-learning/
13. [Online], Kassambara (2018), "Linear Regression Essentials in R", http://www.sthda.com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r/
14. Sarno, R., Dewandono, R. D., Ahmad, T., Naufal, M. F., & Sinaga, F. (2015). "Hybrid Association Rule Learning and Process Mining for Fraud Detection". IAENG International Journal of Computer Science, 42(2), pp. 59-72.
15. Shajideen, Nasreen M and V Bindu. "Spam Filtering: A Comparison Between Different Machine Learning Classifiers." 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) (2018): 1919-1922.
16. W. Peng, L. Huang, J. Jia and E. Ingram, "Enhancing the Naive Bayes Spam Filter Through Intelligent Text Modification Detection," 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), New York, NY, 2018, pp. 849-854. doi: 10.1109/TrustCom/BigDataSE.2018.00122
17. S. M. Bhalerao and M. Dalal, "Improved social network aided personalized spam filtering approach using RBF neural network," 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, 2017, pp. 1-5. doi: 10.1109/I2C2.2017.8321938
18. W. You, K. Qian, D. Lo, P. Bhattacharya, M. Guo and Y. Qian, "Web Service-Enabled Spam Filtering with Naïve Bayes Classification," 2015 IEEE First International Conference on Big Data Computing Service and Applications, Redwood City, CA, 2015, pp. 99-104. doi: 10.1109/BigDataService.2015.19
19. N. M. Shajideen and V. Bindu, "Conventional and Ontology Based Spam Filtering," 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR), Ernakulam, 2018, pp. 1-3. doi: 10.1109/ICETIETR.2018.8529061
20. Godwin Caruana and Maozhen Li. 2008. A survey of emerging approaches to spam filtering. ACM Comput. Surv. 44, 2, Article 9 (March 2008), 27 pages. DOI:https://doi.org/10.1145/2089125.2089129
21. R. Shams and R. E. Mercer, "Classifying spam emails using text and readability features," Proc. - IEEE Int. Conf. Data Mining, ICDM, pp. 657–666, 2013.
22. N. O. F. Elssied, O. Ibrahim, and A. H. Osman, "A novel feature selection based on one-way ANOVA F-test for e-mail spam classification," Res. J. Appl. Sci. Eng. Technol., vol. 7, no. 3, pp. 625–638, 2014.
23. K. Kowsari, et al., "HDLTex: Hierarchical Deep Learning for Text Classification," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, 2017, pp. 364-371. doi: 10.1109/ICMLA.2017.0-134
24. Saha S., DasGupta S., Das S.K. (2019), "Spam Mail Detection Using Data Mining: A Comparative Analysis". In: Satapathy S., Bhateja V., Das S. (eds) Smart Intelligent Computing and Applications. Smart Innovation, Systems and Technologies, vol 104. Springer, Singapore. Doi: 10.1007/978-981-13-1921-1_56
25. Barushka, A., & Hajek, P. (2019). "Review Spam Detection Using Word Embeddings and Deep Neural Networks". Artificial Intelligence Applications and Innovations, 340–350. doi:10.1007/978-3-030-19823-7_28
26. Barushka, A., & Hajek, P. (2018). "Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks". Applied Intelligence. doi:10.1007/s10489-018-1161-y
27. Barushka, A., & Hajek, P. (2018). "Spam Filtering in Social Networks Using Regularized Deep Neural Networks with Ensemble Learning". Artificial Intelligence Applications and Innovations, 38–49. doi:10.1007/978-3-319-92007-8_4
28. Shuaib Bobi, et al. (2019). "Whale optimization algorithm-based email spam feature selection method using rotation forest algorithm for classification". SN Applied Sciences. Doi: 1.10.1007/s42452-019-0394-7.

29. Mirza, N., Patil, B., Mirza, T., & Auti, R. (2017). "Evaluating efficiency of classifier for email spam detector using hybrid feature selection approaches". 2017 International Conference on Intelligent Computing and Control Systems (ICICCS). doi:10.1109/iccons.2017.8250561
30. Habib, M., et al. (2018). "Automatic Email Spam Detection using Genetic Programming with SMOTE". Fifth HCT Information Technology Trends (ITT). doi:10.1109/ctit.2018.8649534
31. Olatunji, S. O. (2017). "*Improved email spam detection model based on support vector machines"* Neural Computing and Applications. doi:10.1007/s00521-017-3100-y
32. Gupta, V., Mehta, A., Goel, A., Dixit, U., & Pandey, A. C. (2018). "*Spam Detection Using Ensemble Learning"*. Advances in Intelligent Systems and Computing, 661–668.doi:10.1007/978-981-13-0761-4_63

AUTHORS PROFILE

Ms. Tarika Verma has passed B.Tech. in 2015 and M.Tech. in 2017 in Computer Science and Engineering from University Institute of Engineering and Technology, Maharshi Dayanand University, Rohtak, India. She had topped in M.Tech (CSE) at UIET, M.D. University in 2017. She has also worked as Assistant Professor at AIJHM College, Rohtak. She is currently pursuing Ph.D. in Computer Science at M. D. University, Rohtak. She has published several research papers in Journals and Conference Proceedings. Her research interests include IoT, Machine Learning, Big Data Analytics and Data Mining.

Dr. Nasib Singh Gill is at present senior most Professor of Dept. of CS & Applications, M. D. University, Rohtak, India and is working in the Dept. since 1990. He earned his Doctorate in CS in the year 1996 and carried out his Post-Doctoral research at Brunel University, West London during 2001-2002. He is a recipient of Commonwealth Fellowship Award of British Government for the Year 2001. Besides, he also has earned his MBA degree. He has published more than 245 research papers in reputed National & International Journals, Conference Proceedings, Bulletins, Edited Books, and Newspapers. He has authored seven books. He is a Senior Member of IACSIT as well as a fellow of several professional bodies including IETE and CSI. He has been serving as Editorial Board Member, Guest Editor, Reviewer of International/National Journals and a Member of Technical Committee of several International/National Conferences. He has guided so far 9 Ph.D. scholars as well as guiding about 7 more scholars presently in the areas – IoT, Machine Learning, Information and Network Security, Computer Networks, Measurement of Component-based Systems, Complexity of Software Systems, Decision Trees, Component-based Testing, Data mining & Data warehousing, and NLP.