www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 04, Issue 02, February 2024, pp : 314-318

e-ISSN : 2583-1062

Impact Factor : 5.725

SPAM IDENTIFICATION

Anil Yadav[1], Abhishek Shukla[2], Aryan Tesia[3], Gautam Shukla[4], Mrs Aparna Majare[5]

[1,2,3,4]Dept - Se-Extc Mumbai University, India.

[5]Assistant Professor, Dept, Extc Mumbai University, India.

DOI: https://www.doi.org/10.58257/IJPREMS32703

ABSRACT

We delve deep into modern content-based e-mail spam filtering methods, focusing on machine learning-driven filters and their adaptations. Our discussion covers key concepts, methods, notable advancements, and recent innovations in this field. Initial analysis reveals the basics of e-mail spam filtering and the role of feature engineering. We conclude by exploring techniques, methodologies, evaluation criteria, and insights from recent progress, paving the way for future research.

Keywords: SVM Classifier, Spam Email Classification, Data Mining, Machine Learning, Deep Learning, etc

1. INTRODUCTION

OVERVIEW: In present times the commercial or bulk e-mails have become a really major problem. Spam nowadays is a waste of storage space, time and bandwidth for communication. From many years the problem caused by spam or fraud mails is increasing. In recent studies, 77% of all mail is spam that comes around a value of 15 billion emails per day and costs Internet users about $ 300 million per year. Today for email filtering, knowledge Engineering and Machine Learning are two most successful approaches. In knowledge engineering approach the hard and fast rule is specifying a set of principles according to which email is classified as spam or ham. Application of this method, doesn't shows any promising results because the rules should be necessary. Constantly updating the rules and methods just causes waste of time and requires more maintenance. As compared to knowledge Engineering, Machine learning is more appropriate approach. It does not have to specify any rules. A set of pre-classified e-mail messages is used here in place of set of rules. Machine learning approaches have a wide range of Importance and a lot of algorithms can be used for e-mail filtering and classification. These include Support Vector Machine, Naïve Bayes

2. LITERATURE REVIEW

1.Inuwa-Dutse (2018) proposed a real-time spam detection system for Twitter, using a combination of ML classifiers, such as Support Vector Machine (SVM), Random Forest (RF), Multi-Layer Perception (MLP), Gradient Boosting, and Maximum Entropy. They used a Honeypot dataset, as well as a manually and automatically annotated spam dataset (SPD) to evaluate their system. They achieved an accuracy of 97.71%, a precision of 99%, a recall of 97%, and an F-score of 98%. However, they noted that their system faced limitations with dealing with lengthy tweets, which could affect the spamming activity detection.

2. Aiyar and Shetty (2018) applied ML models, such as SVM, RF, and Naive Bayes (NB) to detect hate speech in YouTube comments. They used N-grams based features to represent the comments, and obtained an F1-score of 0.97. They suggested that better word representation techniques, such as word embeddings, could improve the performance of their system.

3. Alharthi (2018) worked on sentiment analysis of Arabic tweets, using Long ShortTerm Memory (LSTM) models. They collected over 10,000 tweets via the Twitter API, and annotated them manually with positive, negative, or neutral labels. They achieved an accuracy of 0.97, but noted that the system classification depended on the tweet length, and that shorter tweets were more difficult to classify.

www.ijprems.com
editor@ijprems.com

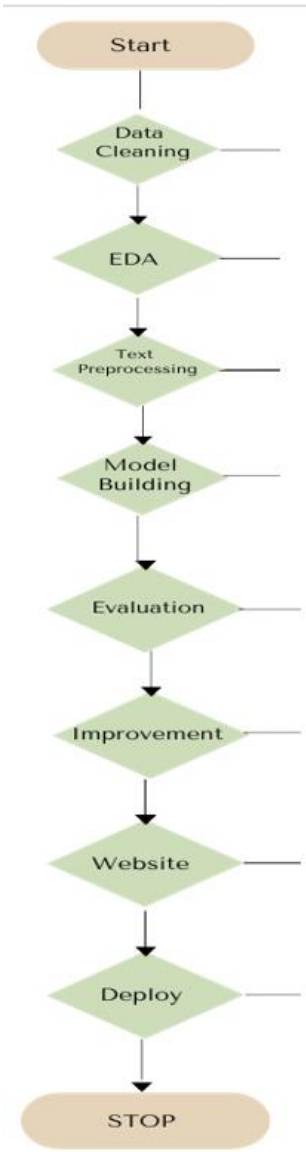
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 04, Issue 02, February 2024, pp : 314-318

e-ISSN : 2583-1062

Impact Factor : 5.725

3. FLOWCHART



SOFTWARE USED:

1.COMMAND PROMPT(CMD):CMD is a command-line interpreter in Windows OS, used for executing commands and scripts. It can be used to navigate files, run certain programs, or perform basic tasks related to spam filtering configurations or management.

2.Jupyter Notebook: Jupyter Notebook is an open-source web application that allows creating and sharing documents containing live code, equations, visualizations, and narrative text. It's used in spam identification for developing and testing machine learning models or algorithms.

3.Visual Studio Code: VS Code is a lightweight but powerful source-code editor with support for various programming languages. It's used in spam identification for writing, editing, and debugging code related to spam filtering algorithms or scripts.

4.Python 3: Python 3 is a widely used high-level programming language known for its simplicity and readability. In spam identification, Python 3 is used for developing machine learning models, data analysis, and scripting tasks for spam detection systems.

5.PyCharm:PyCharm is an integrated development environment (IDE) for Python development. It provides tools for coding assistance, debugging, and intelligent code analysis, aiding developers working on spam identification algorithms or scripts.

6.Streamlit:Streamlit is an open-source Python library used for building web applications for data science and machine learning projects. In spam identification, Streamlit can be utilized to create user-friendly interfaces for displaying spam identification results or interacting with spam detection models

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

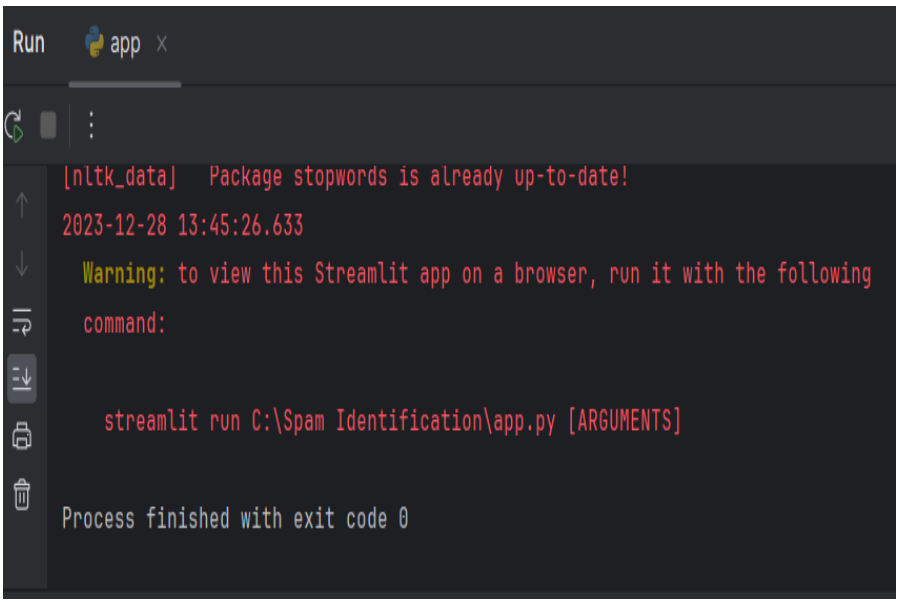
Vol. 04, Issue 02, February 2024, pp : 314-318

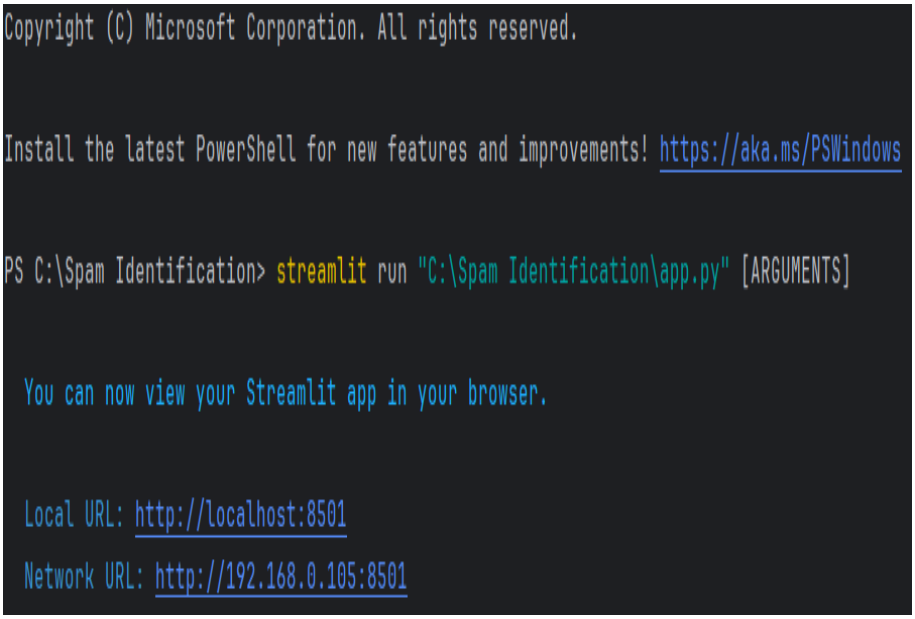
e-ISSN : 2583-1062

Impact Factor : 5.725

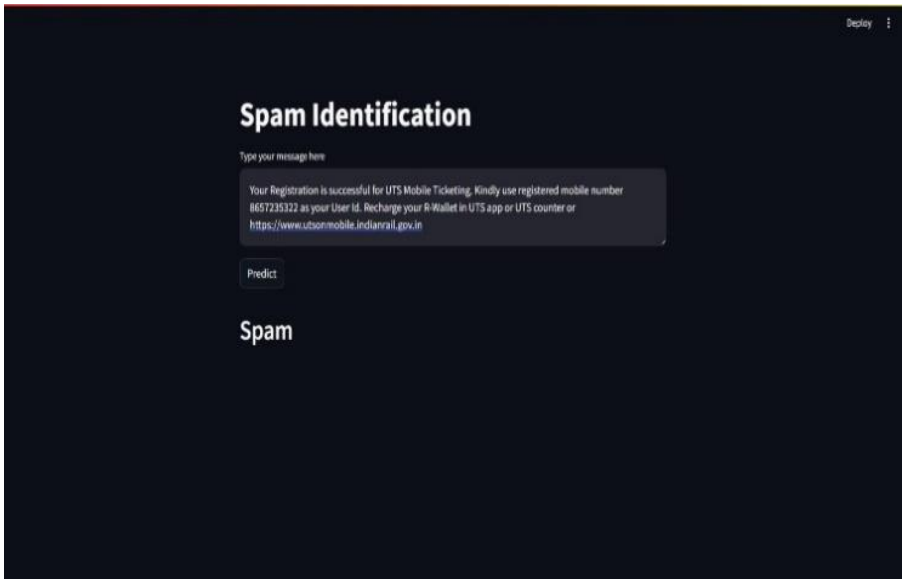
4. OUTPUT AND RESULT

1. URL & STREAMLIT RUN:-





TESTING:



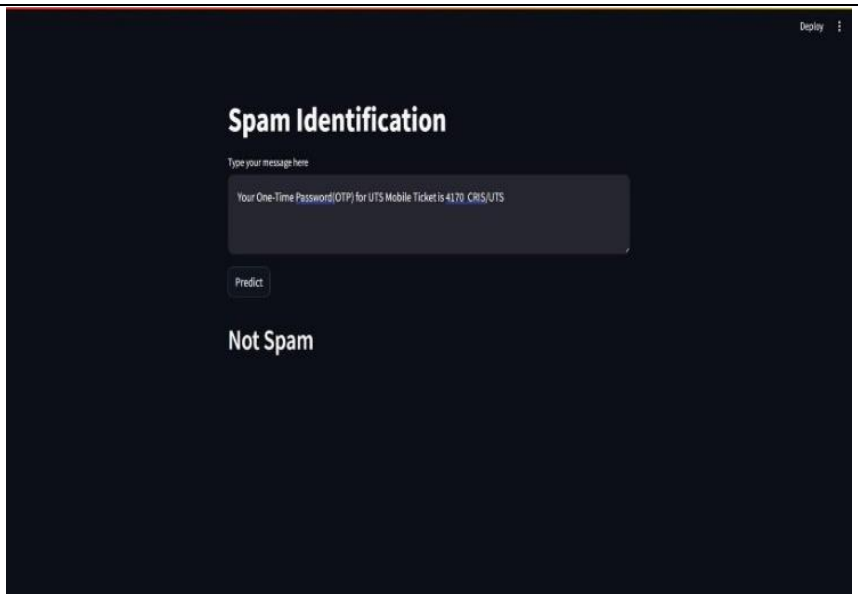
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 314-318

e-ISSN : 2583-1062

Impact Factor : 5.725



5. ADVANTAGE & DISADVANTAGE

ADVANTAGE:

1.Effectively filters unwanted emails

2.Enhances cybersecurity

3.Protects personal information

4.Reduces malware and phishing threats

5.Increases user productivity

DISADVANTAGE:

1.Possibility of false positives

2.Evolving tactics of spammers

3.Over-filtering legitimate emails

4.Resource-intensive for large-scale operations

5.Language and cultural barriers

6. CONCLUSION

Email has been the most important medium of communication nowadays, through internet connectivity any message can be delivered to all aver the world. More than 270 billion emails are exchanged daily, about 57% of these are just spam emails. Spam emails, also known as non-self, are undesired commercial or malicious emails, which affects or hacks personal information like bank ,related to money or anything that causes destruction to single individual or a corporation or a group of people. Besides advertising, these may contain links to phishing or malware hosting websites set up to steal confidential information. Spam is a serious issue that is not just annoying to the end-users but also financially damaging and a security risk. Hence this system is designed in such a way that it detects unsolicited and unwanted emails and prevents them hence helping in reducing the spam message which would be of great benefit to individuals as well as to the company .In the future this system can be implemented by using different algorithms and also more features can be added to the existing system.

7. REFERENCE

[1] Shukor Bin Abd Razak, Ahmad Fahrulrazie Bin Mohamad “Identification of Spam Email Based on Information from Email Header” 13th International Conference on Intelligent Systems Design and Applications (ISDA), 2013

[2] Trivedi, Shrawan Kumar. "A study of machine learning classifiers for spam detection." Computational and Business Intelligence (ISCBI), 2016 4th International Symposium on. IEEE, 2016. [10] You, Wanqing, et al. "Web Service-Enabled Spam Filtering with Naïve Bayes Classification." 2015 IEEE First International Conference on Big Data Computing Service and Applications (BigDataService). IEEE, 2015.

[3] Sahın, Esra, Murat Aydos, and Fatih Orhan. "Spam/ham e-mail classification using machine learning methods based on bag of words technique." 2018 26th Signal Processing and Communications Applications Conference (SIU). IEEE, 2018

[4] Mujtaba, Ghulam, et al. "Email classification research trends: Review and open issues." IEEE Access 5 (2017).

[5] Ravinder Kamboj, "A rule based approach for spam detection" ,Computer Science and Engineering Department, Thapar University, India, July 2010.

[6] M. U. S. Khan, M. Ali, A. Abbas, S. U. Khan, and A. Y. Zomaya, "Segregating spammers and unsolicited bloggers from genuine experts on Twitter," IEEE Trans. Dependable Secure Comput., vol. 15, no. 4, pp. 551–560, Jul./Aug. 2018.

[7] E. Blanzieri and A. Bryl, E-mail Spam Filtering with Local SVM Classifiers, University of Trento, Trento, Italy, 2008

[8] W. N. Gansterer, A. G. K. Janecek, and R. Neumayer, "Spam filtering based on latent semantic indexing," in Survey of Text Mining II, pp. 165–183, Springer, New York, NY, USA, 2008.

[9] A. Bhowmick and S. M. Hazarika, "Machine learning for E-mail spam filtering: review, techniques and trends," 2016, (1) (PDF) Machine Learning for E-mail Spam Filtering: Review,Techniques and Trends (researchgate.net)

[10] E. P. Sanz, J. M. Gómez Hidalgo, and J. C. Cortizo Pérez, "Chapter 3 email spam filtering," Advances in Computers, vol. 74, pp. 45–114, 2008.

[11] H. Takhmiri and A. Haroonabadi, "Identifying valid email spam emails using decision tree," International Journal of Computer Applications Technology and Research, vol. 5, 2016.

[12] P. S. Keila and D. B. Skillicorn, "Structure in the Enron email dataset," Computational & Mathematical Organization Theory, vol. 11, no. 3, pp. 183–199, 2005.

[13] Q. Wang, Y. Guan, and X. Wang, "SVM-based spam filter with active and online learning," in Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006, NIST, Gaithersburg, MD, USA, November 2006

[14] W. Peng, L. Huang, J. Jia, and E. Ingram, "Enhancing the naive bayes spam filter through intelligent text modification detection," in Proceedings of the 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), IEEE, New York, NY, USA, August 2018.

[15] K. Tretyakov, "Machine learning techniques in spam filtering," in Data Mining Problem-Oriented Seminar, vol. 3, no. 177, pp. 60–79, 2004.

[16] A. K. Sharma and S. Sahni, "A comparative study of classification algorithms for spam email data analysis," International Journal on Computer Science and Engineering, vol. 3, no. 5, pp. 1890–1895, 2011

[17] H. Xu, W. Sun, and A. Javaid, "Efficient spam detection across online social networks," in Proceedings of the 2016 IEEE International Conference on Big Data Analysis (ICBDA), IEEE, Hangzhou, China, March 2016.

[18] A. H. Wang, "Detecting spam bots in online social networking sites: a machine learning approach," in Proceedings of the IFIP Annual Conference on Data and Applications Security and Privacy, Springer, Berlin, Germany, June 2010.