IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SMART SPAM DETECTION SYSTEM ON SOCIAL MEDIA PLATFORM LIKE YOUTUBE USING MACHINE LEARNING

NAUMAN ZARI, SANKET POTGHAN, SWEETY KALE.

Dept. Of Computer Engineering, Sinhgad Institute Of Technology, Lonavala, Pune, Maharashtra, India - 410401

Abstract: The popularity of Google's video distribution platform, YouTube, has led to an influx of users. Unfortunately, this success has also attracted a significant number of malicious individuals who aim to promote their own videos or distribute viruses and malware. Due to YouTube's limited comment moderation tools, spam comments increase rapidly, often leading video owners to disable the comment section. Detecting and classifying spam comments pose a challenge as they are usually short, filled with slangs, symbols, and abbreviations.

In this research paper, we conducted an evaluation of various high-performance classification techniques to detect and analyze spam comments. Through statistical analysis, we found that Decision trees, Logistic regression, Bernoulli Naive Bayes, Random Forests, Linear, Gaussian, and SVM classifiers are all statistically equivalent in terms of their maximum performance rate, with a confidence level of 99.9%. Hence, it is crucial to find a way to identify and report these comments before innocent users encounter them.

To address this issue, we propose utilizing Spyder, a popular data analysis tool, to analyze the data and create well-fitted models for the project. These models can then be used to identify patterns and make predictions, which can be integrated into further algorithms to obtain the desired output.

Overall, this paper emphasizes the importance of detecting spam comments on YouTube and presents an evaluation of multiple classification techniques as potential solutions.

Index Terms **– Spam Detection; SVM; Naïve Bayes; Binary Classification; Logistic Regression**

I. INTRODUCTION

In recent years, informal online communities such as Facebook and YouTube have become increasingly common platforms in people's daily lives. These social media platforms are used by individuals to stay connected with friends and family, as well as to share thoughts and ideas through blogs. However, due to their popularity, these platforms attract a large number of users and also become easy targets for spammers. For instance, there has been a rise in the number of makeup tutorial videos on YouTube, created by bloggers known as "beauty gurus" or "beauty influencers," with a significant audience comprised of teenage girls. Currently, YouTube sees the generation of 400 million new videos each day, thanks to its extensive user base of 200 million. This vast environment provided by YouTube also presents an opportunity for spammers to create and distribute irrelevant content.

These spammers aim to deceive users by enticing them to click on links that lead to malicious websites containing malware, phishing attempts, or scams. This project aims to predict spam comments present in the comment sections of YouTube videos using machine learning techniques, a subset of artificial intelligence. The chosen approach is supervised learning, which relies on a large number of labeled datasets.

The proposed classification algorithm for this project is logistic regression. By utilizing machine learning, the project seeks to introduce the concept and outline the prediction technique, highlighting its superiority over conventional data analysis methods and its potential to improve prediction accuracy.

Spam comments are typically entirely unrelated to the video content and are often generated by automated bots posing as users. The comments section becomes a target for spammers to post irrelevant messages, comments, links, and ideas. Machine learning is the method used to extract, transform, load, and predict relevant information from vast datasets. It helps identify patterns and structures the data in a comprehensible format for further utilization.

Classification and prediction are two types of data analysis used to categorize fundamental data classes and forecast future trends. The presence of malicious spam comments can tarnish the positive aspects of the videos they are posted on. While the project to predict spam comments is currently underway, an accurate prediction model has not yet been finalized and established.

II. LITERATURE SURVEY

Paper name:- Network based spam filter on YouTube

Methodology: -Bayesian classifier

Publication year: - 2020

Conclusion: This paper thoroughly examined various technologies for identifying and blocking spam emails and proposed a model based on a naive Bayesian classifier to determine whether an email is spam or not. The experimental results demonstrated the effectiveness of the model in a large spam dataset. While the naive Bayesian model theoretically boasts the lowest error rate among classification methods, it is essential to acknowledge the impractical assumptions made by this model regarding attributes.

Paper name :-Analysis and detection of spam accounts in social networks

Methodology :- Extreme machine learning algorithms

Publication year :- 2020

Conclusion: Several approaches have been developed to address the issue of identifying Twitter spam. Many of these methods employ machine learning algorithms to differentiate between spam and non-spam content [1]. Initial research efforts have explored various factors, including account and content characteristics, such as account age, follower/following counts, URL ratio, and tweet length, to differentiate between spammers and legitimate users. Although these features can be extracted efficiently, they are also susceptible to manipulation. In summary, extensive research has been conducted on machine learning algorithms for detecting Twitter spam.

Paper name: - A neural network-based ensemble approach for spam detector

Methodology: - CNN architecture

Publication year: - 2020

Conclusion: In recent times, deep learning techniques have demonstrated promising outcomes in various natural language processing tasks. We aim to leverage the advantages offered by these methods for our specific problem. With this objective in mind, we present an ensemble approach for tweet-level spam detection. Our approach involves the utilization of convolutional neural networks (CNNs) as the basis for multiple deep learning models. The ensemble comprises five distinct CNNs, along with a feature-based model. Each CNN incorporates different word embeddings, such as GloVe and Word2Vec, during the training process.

Paper name :- Design of machine learning approach for spam tweet detection.

Methodology :- Naive Bayesian , SVM classifiers.

Publication year :- 2019

Conclusion: Naïve Bayes is a supervised machine learning algorithm utilized for classification tasks, relying on Bayes' theorem. The term "naïve" stems from its assumption of conditional independence among predictors, considering all features within a class as unrelated. It serves as a supervised machine learning algorithm capable of performing regression and classification tasks. Support Vector Machines (SVM) plot data points in an n-dimensional space, where 'n' represents the number of features, and classification is achieved by identifying an appropriate hyperplane that distinguishes between the two classes. In the context of spam message detection, a classifier-based approach is employed.

III. METHODOLOGY AND STRATEGY

Fetching contents and attributes for dataset:

A lot of data is being generated per second and is difficult to manage them and organize it. We, specifically have gone through a lot of You Tube video comment section and collected a dataset which include all kind of sentimental expressions in the form of text. These datasets will play a major role in initiating the learning process.

Storing comments into table:

We have arranged and organized the data collected from various vlogs for the ease of implementation and representation of comments.

Implementation Strategy:

1. Create single training data set file using the 2 of 3 YouTube data files.
2. Since the data set is text data, need to obtain the features from the comment(CONTENT) field.
3. From the data set only "CONTENT" field in the file is relevant to obtain the features.
4. Remove the unnecessary special characters (ï»¿) from the comment field.
5. Create SPAM dictionary of words which usually appear in Spam messages by observing Spam class samples.
6. Count the number of spam words in the comment using the spam dictionary .
7. To check if the comment contains strings "http“ ,"www" or ".com" string which represent promotions and could be SPAM and set IS_HTTP=1 else .
8. To calculate the ratio of Spam words to number of words in comment, first remove the English ‘STOPWORDS’ like (I ,me ,the ,etc.) from the comment field.
9. Get the length of the comment as long length comments are usually spam comments. Count the number of words from the comment after removing the stop words.
10. Calculate the ratio of Spam words and total number of words in comment.
11. Execute the Naïve Bayes classifier on test data and validate it with the test data.

IV. PROJECT RESOURCES

The resources required for a spam comment detection project will depend on the scope, complexity, and methodology chosen for the project. Here are some common resources that may be required:

1. Data sources: Trends on social media platforms, comment section and other relevant data sources are needed to develop and validate the spam comment detection classification algorithm. Depending on the project requirements, these data sources can be obtained from various social media platform.

2. Hardware infrastructure: The hardware requirements will depend on the size and complexity of the problem. A powerful computer or server with sufficient memory and processing power may be necessary to handle large-scale classification and distinguishing problems.

3. Software tools and libraries: A variety of software tools and libraries can be used to develop the algorithm. These may include programming languages like Python libraries like seaborn, matplotlib and nltk.corpus etc.

4. Visualization: Visualization is essential to make the spam comments detection solution/outcomes accessible to end-users.

In summary, the resources required for a shortest route optimization project can be substantial, and it's important to carefully plan and manage these resources to ensure the success of the project.

V. TOOLS AND TECHNOLOGIES USED

Gaussian Naïve Bayes - Naïve Bayes is a probabilistic machine learning algorithm used for many classification functions and is based on the Bayes theorem. Gaussian Naïve Bayes is the extension of naïve Bayes.

Binary Classification – Binary classification is the task of classifying the elements of a set into two groups (each called class) on the basis of a classification rule.

Nltk.corpus – NLTK corpus readers. The modules in this package provide functions that can be used to read corpus files in a variety of formats.

Seaborn – Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

Technologies – Supervised Classification

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called as targets/labels or categories.

In classification algorithm, a discrete output function(y) is mapped to input variable(x).

$y = f(x)$, where y = categorical output

Algorithm Detail

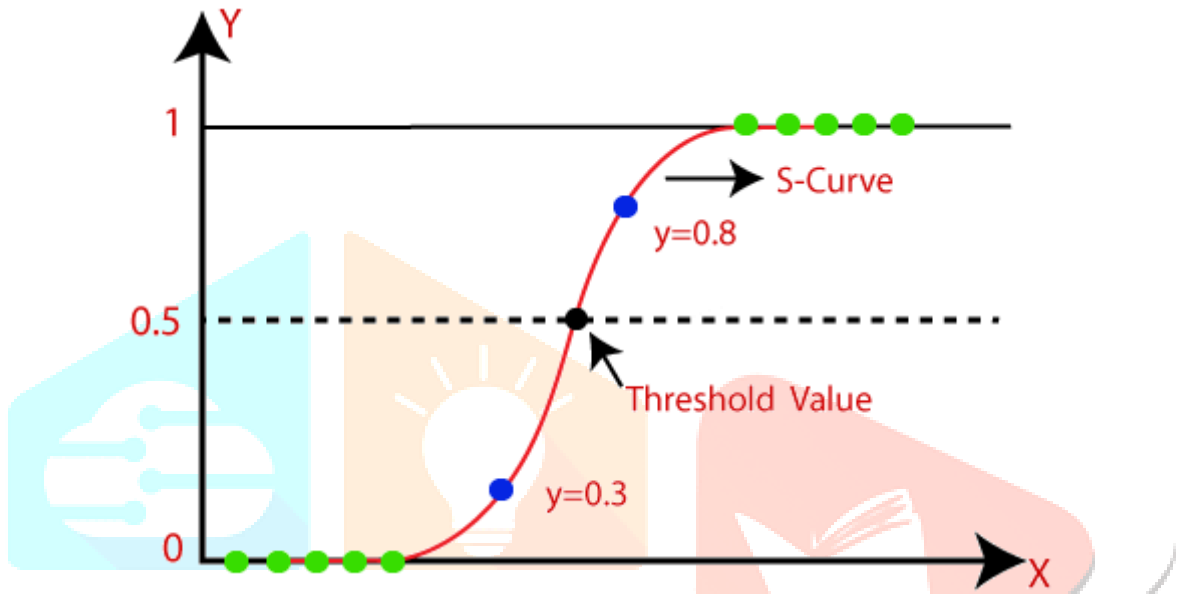
Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

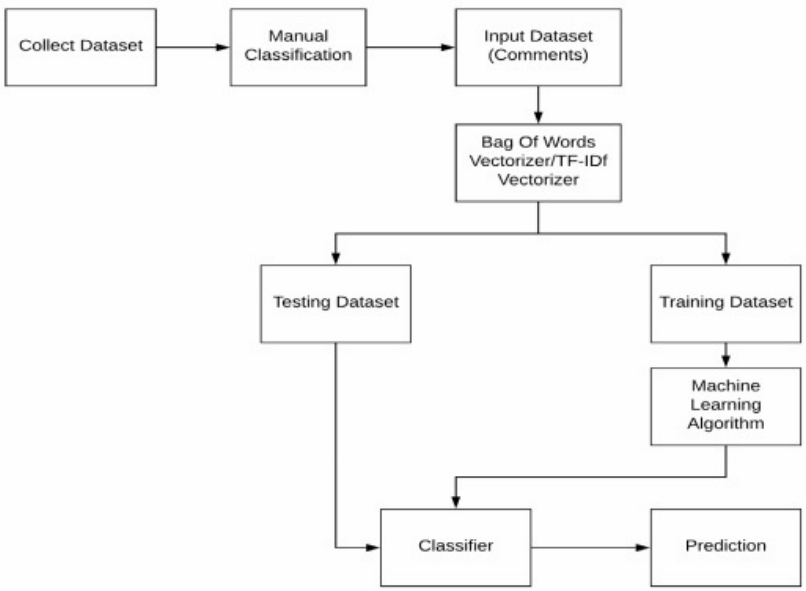
Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

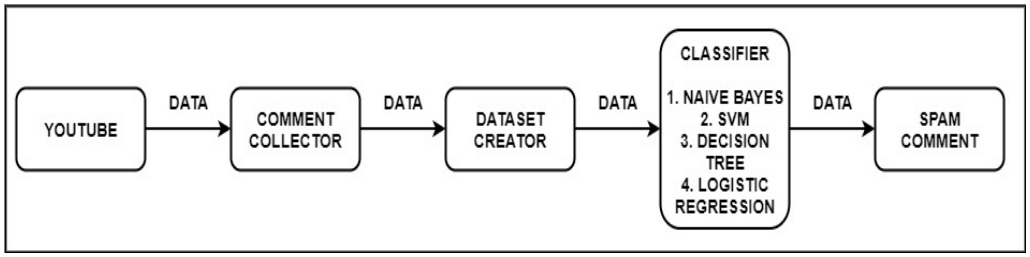


VI. Proposed System Design

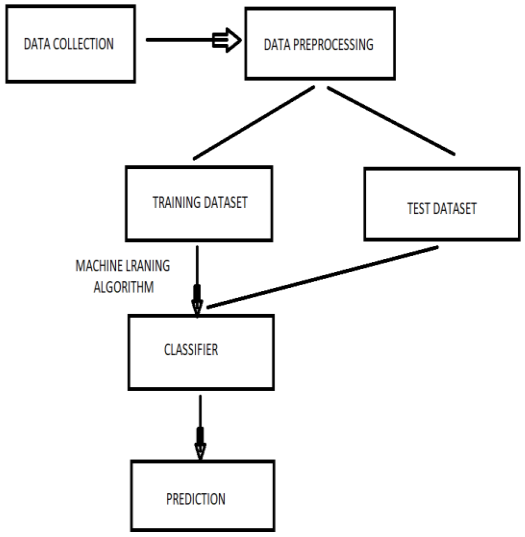
We attempt to detect spam comments by applying conventional machine learning algorithms Naive Bayes along with certain custom heuristics such as N-Grams which have proven to be very effective in detecting and subsequently combating spam comments. We have collected and created three databases composed by real, public and non-encoded data directly extracted from YouTube. We have selected five of the ten most viewed YouTube videos. Each sample represents a text comment posted in the comments section of each selected video. No preprocessing technique was performed. Subsequently, each sample was manually labeled as spam or legitimate (not spam), using a collaborative tagging tool developed for this purpose, called Labeling. The samples have associated a metadata information, such as the author's name and publication date, which have been preserved.



VII. DIAGRAMS



DFD

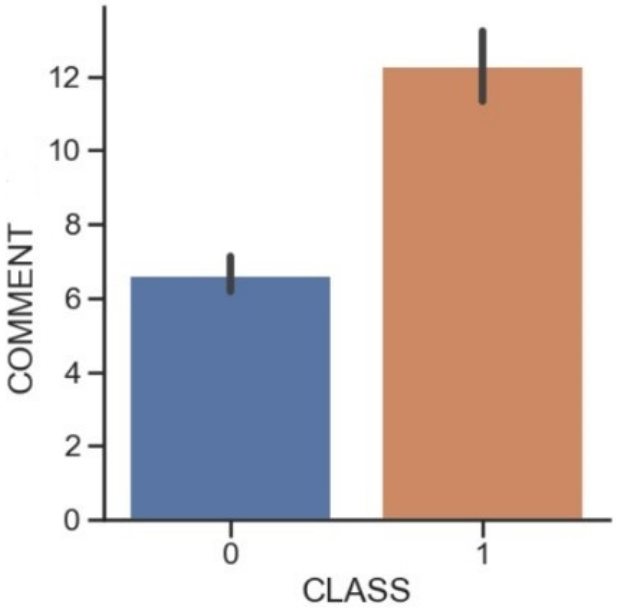


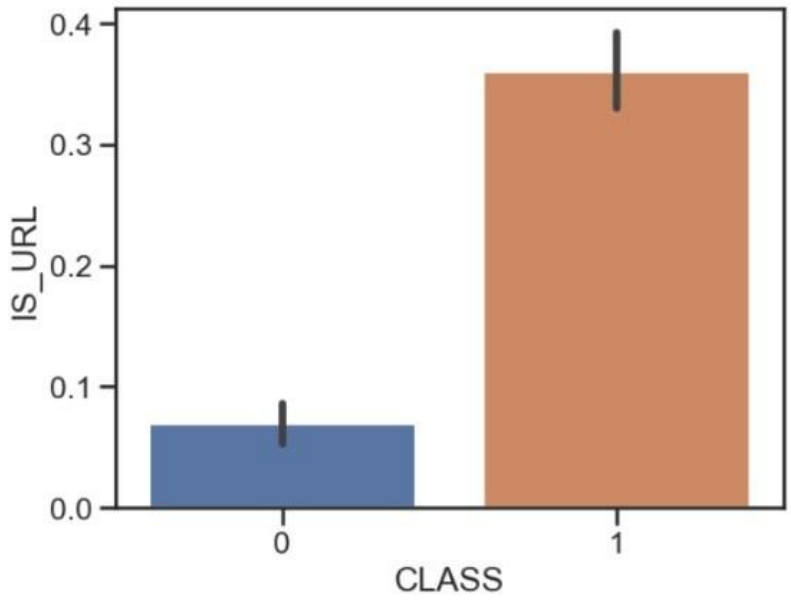
UML

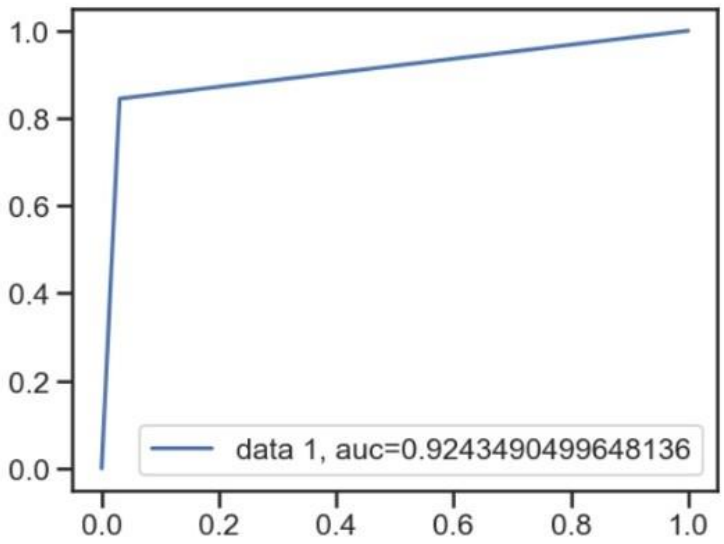
VIII. RESULTS

Social media networks have become popular and this creates the opportunity for the spammers to publish unwanted comments.

Previously, some machine learning algorithms were used for this detection. In the proposed system we also use the advanced machine learning algorithms with advanced features also compares the efficiency of various algorithms by applying them We construct features based on the features obtained from the user profile and the content that they shared. Based on the experiments conducted, it can be expected that existing classifiers widely used in the data mining community can utilize these functions to detect spammers.


COMMENT
12
10
8
6
4
2
0
0
1
CLASS


IS_URL
0.4
0.3
0.2
0.1
0.0
0
1
CLASS


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
data 1, auc=0.9243490499648136

REFERENCES

[1] N. M. Samsudin, C. F. B. Mohd Foozy, N. Alias, P. Shamala, N. F. Othman and W. I. S. Wan Din, "Youtube spam detection framework using Naïve Bayes and logistic regression", *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1508, Jun. 2019.
[2] R. K. Das, S. S. Dash, K. Das and M. Panda, "Detection of spam in Youtube comments using different classifiers", *Advanced Computing and Intelligent Engineering*, pp. 201-214, 2020.
[3] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan and S. A. Razak, "Malicious accounts: Dark of the social networks", Elsevier, 2017, pp. 41-67.
[4] S. Garg Makkar, N. Kumar, M. S. Hossain, A. Ghoneim and M. Alrashoud, "An Efficient Spam Detection Technique Using Machine Learning", *IEEE Transactions on Industrial Informatics*, vol. 17, no. 2, pp. 903-912, Feb. 2021.
[5] C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou and G. Min, "Statistical Features- Based Real- Time Detection of Drifted Twitter Spam", IEEE Transactions, April 2017, pp.914-925.
[6] M. Verma, Divya, S. Sofat, "Techniques to Detect Spammers in Twitter – A Survey", International Journal of Computer Applications, January 2014, Vol. 85, No. 10, pp. 27-32.
[7] A. Gupta and R. Kaushal, "Improving Spam Detection in Online Social Networks", IEEE, 2015.

BIOGRAPHY

*Nauman J Zari
An Undergraduate Scholar pursuing Bachelors of Engineering in Computer from Sinhgad Institute Of Technology

*Sanket A Potghan
An Undergraduate Scholar pursuing Bachelors of Engineering in Computer from Sinhgad Institute Of Technology

*Sweety Kale
An Undergraduate Scholar pursuing Bachelors of Engineering in Computer from Sinhgad Institute Of Technology