© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 14 Issue: 02 | April - June 2023
Paper is available at http://www.jrps.in | Email : info@jrps.in
Refereed & Peer Reviewed

SPAM DETECTION ON SOCIAL MEDIA USING HYBRID ALGORITHM

Raghuveer Pahade[1], Dr. Pankaj Richariya[2]
[1]Research Scholar, Department of CSE, BITS, Bhopal,
[2]HOD, Department of CSE, BITS, Bhopal

Abstract

This work presents a comprehensive study on the detection of YouTube Spam comments. The study was aimed at investigating the impact of using machine learning algorithms on the accuracy of detecting spam comments. The study used a dataset of YouTube comments collected from Kaggle sources and underwent a pre-processing stage to ensure the data was in a format suitable for analysis. Three machine learning algorithms were used to build models for the classification of YouTube comments as spam or not, these algorithms include Logistic Regression, Random Forest, and Ada Boost Classifier. Additionally, a hybrid model was developed by combining the best-performing base models, Random Forest, Logistic Regression, and Ada Boost Classifier using a voting classifier.

The models were evaluated using three evaluation metrics: accuracy, precision, and recall. The results showed that the hybrid model outperformed the other models, achieving an accuracy of 97.8%, precision of 99.9% and recall of 95.9%. The study also compares the results with existing work and found that the proposed hybrid model achieved higher precision, accuracy, and recall. The study concludes that the use of a hybrid model is a suitable solution for detecting YouTube spam comments and provides promising results.

However, the study also acknowledges that there are limitations in the study, including the limited size of the dataset and the limitations of machine learning algorithms used. Future work includes exploring alternative machine learning algorithms and increasing the size of the dataset to further improve an accuracy of the spam detection. Overall, the results of this study have implications for the development of more robust and the effective spam detection systems for YouTube comments.

Keyword: Spam Detection, Logistic Regression, Random Forest, Ada Boost Classifier, Hybrid Algorithm

1. Introduction

Spam detection on social media platforms has become an increasingly important issue in recent years. Social media has transformed the way people communicate and share information, but it has also become a platform for the spread of spam. This spam can take many forms, including commercial promotions, phishing scams, and fake news. Spam not only affects the user experience but can also pose a threat to the security and privacy of users.

Spam on social media platforms has become a growing problem in recent years. Social media platforms, like Twitter, Facebook, and YouTube, have millions of users who generate a vast amount of content, including comments, posts, and messages. YouTube was launched in 2005 and is a website where users may upload and share videos. More than 2 billion people use it every month, making it one of the most visited websites ever. Users can upload, share, and view videos on a wide range of topics, including music, entertainment, education, and news. In addition to videos, YouTube also allows users to comment on videos and interact with one another through its social media-like features. The platform has become an important source of information, entertainment, and communication for people all over the world and has had a profound impact on the way we consume and share media. Despite its popularity and importance, YouTube also faces a number of challenges, including the proliferation of spam and inappropriate content, which has raised concerns about the quality and reliability of the information shared on the platform. However, this increase in user-generated content has also led to the proliferation of spam.

There are various approaches that have been proposed for detecting spam on YouTube, including rule-based methods, machine learning techniques, and hybrid approaches that combine multiple methods.

- **Rule-based approach:** This approach uses a set of predefined rules and heuristics to identify and flag comments that are likely to be spam. The rules may include checking for specific keywords, patterns in the text, or the presence of links.

- **Machine learning approach:** This approach uses several machine learning algorithms, like Naive Bayes, decision trees, or Support Vector Machines, to train a model on a dataset of YouTube comments and then use the trained model to identify new comments as spam or not spam.

- **Deep learning approach:** This approach uses deep neural networks to learn complex representations of the data and make predictions about the class of a comment as spam or not spam.

- **Hybrid approach:** This approach combines multiple methods, such as rule-based, machine learning, and deep learning, to take advantage of the strengths of each method and improve the overall performance of the spam detection system.

- **Natural Language Processing (NLP) approach:** This approach uses NLP techniques, such as text classification, sentiment analysis, and text clustering, to analyze the text of comments and identify patterns that are indicative of spam.

Each of these approaches has its own strengths and limitations, and the choice of approach will depend on the specific requirements and goals of the spam detection system. In recent years, machine learning as well as deep learning approaches have become increasingly popular for detecting spam on YouTube, as these methods can be trained on large datasets and can adapt to changing patterns in the data.

With raising popularity of social media platforms, such as YouTube, the amount of user-generated content has also increased dramatically. However, this increase has also led to the proliferation of spam comments, which can be annoying, disruptive, and even harmful to users. Spam comments on YouTube can range from commercial promotions to phishing attempts and fake news, and can negatively impact the user experience and undermine the trust in the platform.

Therefore, there is a need for effective methods to detect and prevent spam on YouTube. The goal of this thesis is to investigate and develop a robust spam detection system for YouTube comments that can accurately identify and flag spam comments. By doing so, this thesis aims to improve the user experience on the YouTube and help maintain the integrity of the platform.

2. Literature Review

[1] Many classification algorithms were constructed by the author in order to distinguish spam comments from genuine ones on YouTube videos; their performance metrics were analysed, and the advantages of using an ensemble classifier over a single classifier technique were emphasised. Nowadays, several malicious attacks have been made against online social networks. Although giving us a place to voice our opinions freely, the general public is in risk because of the abuse of this potent tool. The author of this research has been sifting through YouTube comment data in an effort to identify and eliminate spam comments. The effectiveness of Bagging was measured against that of other state-of-the-art text categorization algorithms like Naive Bayes (the ensemble classifier). Ensemble classifiers have been shown to perform better than individual ones in most situations.

[2] The Author has conducted an analysis of numerous high-performing categorization methods for this same aim in this study. Decision trees, Bernoulli Naive Bayes, logistic regression, random forests, linear and Gaussian support vector machines are statistically equal, according to the findings of the investigation. From this, the author has developed TubeSpam, a reliable web-based method for filtering YouTube comments. The majority of the tested categorization algorithms are recommended for use in the YouTube's comment spam filter. In fact, majority of them had blocked ham rates of 0% or less and accuracy rates of 90percentage points or higher. The Nemenyi post hoc test was used to assess the approaches against one another after the Friedman test had shown that the findings were not the product of random chance. The posthoc analysis demonstrated a 99.9% confidence level that the performances of CART, LR, NB-B, RF, SVM-L, and SVM-R are statistically equal.

[3] Methods for identifying hallmarks of spam in the video-sharing comments are proposed here. In this study, gathering relevant datasets is the starting point of the process. The data utilised in this study comes from the UCI Machine Learning repository. The next step is the creation of a framework and the beginning of experimentation. Tokenization and lemmatization are pre-processing steps that will be applied to the dataset. Next, six classifiers—Random Tree, Random Forest, Naive KStar, Decision Table, Bayes, and Decision Stump—were put through their paces in studies designed to determine the best method for detecting spam. The maximum rate of accuracy achieved in this study was 90.57%, while the lowest was 58.86%.

[4] Author presents a novel approach to identifying spam accounts based on their behaviour using extreme learning machine (ELM), a supervised machine learning technique. They begin their research by scraping Sina Weibo for messages. To further improve the efficacy of ELM-based spam accounts detection algorithm, the author has chosen to use characteristics gleaned from social interactions, message contents,and user profile data. Lastly, the author conducts experimental and evaluative research to confirm the detectability of the spam accounts. As compared to current supervised machine learning approaches, their suggested methodology has the potential to provide superior system function outcomes more quickly.

[5] This work investigates the use of the ensemble classifiers in the construction of a spam categorization model. The purpose of this research is to develop a highly accurate & the sensitive classification model for distinguishing between ham emails and spam emails. In order to find useful characteristics in the Enron email collection, the greedy stepwise feature search technique has been used. Several machine learning classifiers (including Bayesian, Naive Bayes, SVM (support vector machine), J48 (decision tree), Bayesian with the Adaboost, and Naive Bayes with Adaboost) have been compared. A variety of metrics (including False Positive Rate, F-measure (accuracy), and training time) are used to assess the performance of the classifiers in question. SVM has been determined to be the most effective classifier after careful consideration of all of these factors. The proportion of false positives is also rather small. Although it's true that SVM takes a while to train before it can produce a usable model, this is less of a concern when the results on the other parameters are good.

[6] Detecting spam in SMS service is the focus of this research article. As of 2012, up to 30% of SMS texts in certain regions of Asia were spam. Existing email filtering algorithms may perform poorly due to a lack of genuine databases for SMS spams, the short duration of messages, the restricted features, and the casual language used in them. In this study, use a variety of the machine learning algorithms on a database including actual SMS Spams obtained from UCI Machine Learning repository. In the end, compare the outcomes and provide the most effective algorithm for the spam filtering through SMS. The best classifier in this work decreases the total error rate of best model in the original research mentioning this dataset by more than half, as shown by final simulation results using the 10-fold cross validation. Several classification models were used on the SMS Spam dataset, and the results are shown here. Based on simulation data, some of the most effective classifiers for the SMS spam detection are the multinomial naive Bayes with the laplace smoothing and SVM with the linear kernel. Using SVM as the learning technique, the top classifier in an original study mentioning this dataset has an overall accuracy of 97.64%. They find that enhanced naive Bayes, the next best classifier in their studies, achieves an accuracy of 97.50%.

3. Methodology

The methodology of this research outlines the approach taken to detect spam comments on YouTube. The goal of the study was to develop a solution that could accurately classify comments as either spam or not spam, with the aim of improving the quality of information shared on the platform.

The study made use of a machine learning approach, utilizing a dataset from various YouTube videos obtained from Kaggle.com. A range of machine learning algorithms were tested and evaluated, and a hybrid model was ultimately selected as the best-performing solution.

The work that was done to achieve the objective of detecting spam comments on YouTube. It highlights the significance of the study and the importance of reducing the prevalence of spam on the platform. The results of this research provide valuable insights into the problem of spam detection, and offer promising solutions for future work in this area.

Spam detection on YouTube is an important problem, as spam comments can spread misinformation, disrupt communication, and hinder the use of the platform for legitimate purposes. In this thesis, I propose to develop a machine learning model for detecting spam on YouTube using a dataset of the YouTube comments. To do this, I will follow the following steps:

1. **Data collection**: The first step of the proposed work will involve collecting a dataset of YouTube comments for use in training and testing the spam detection model.

2. **Data preprocessing**: Once the data has been collected, the next step will be to preprocess the data to prepare it for analysis. This may involve cleaning and filtering the data, performing word vectorization.

3. **Model development**: The next step will be to develop a spam detection model using machine learning techniques. This may involve selecting an appropriate algorithm, such as a random forest or a support vector machine, and training and evaluating the model using the preprocessed data.

4. **Model evaluation**: After the model has been developed, the next step will be to evaluate its performance using metrics such as accuracy and precision. It is also consider comparing the performance of the model to other models or to a baseline, to assess its effectiveness.

5. **Model fine-tuning**: If the model's performance is not satisfactory, the next step will be to fine-tune the model by adjusting its parameters or preprocessing the data differently. This may involve iterating through steps 3 and 4 until the model's performance is satisfactory.

6. **Model testing**: Once the model's performance is satisfactory, the final step will be to test the model using the test data to ensure that it is generalizing well to unseen data.

By following these steps, I hope to develop a machine learning model that is able to accurately detect spam on YouTube and contribute to the understanding of this important problem.

Model Building & Training

In the training phase, the models were trained on the pre-processed data, and their performance was evaluated depending on several performance metrics, for example accuracy, precision and recall. After evaluating the performance of all the models, the hybrid model, which was a combination of the best-performing base models, was found to have the highest accuracy of 97.8%. This was considered to be the final model, and it was used to classify the YouTube comments as spam or not spam.

Hyper Parameters of the model –

Table 1 Train Test Split Ratio

Train size	80%
Test size	20%

Table 2 Model Hyper parameters

Model	Parameters

CountVectorizer	ngram_range=(1, 1),
	analyzer='word',
	max_features = 2106,
	vocabulary=None,
	lowercase='True',
	binary=False

Logistic Regression	Solver = 'newton-cg',
	Penalty = 'l2'

Random Forest	n_estimators = 70,
	random_state = 42

AdaBoost	n_estimators = 50,
	random_state = 28

Voting Classifier	Estimators = 3 Base Classifiers,
	Voting = 'soft',
	n_jobs = -1

4. Results

The results obtained from the evaluation of the models used in this thesis for the classification of YouTube comments as spam or not spam, are as follows:

Logistic Regression Model: The accuracy of this model was found to be 95%. In terms of precision, the model achieved a score of 99.9% which indicates the proportion of true positive predictions among the positive predictions made. Additionally, the recall score of the Logistic Regression model was 90.5%, which reflects the proportion of the positive instances that were correctly identified by a model.

Random Forest Model: This Random Forest model has a 96.4% success rate. The model's accuracy was 99.9 percent, indicating that 99.9 percent of the time, the model generated accurate predictions. This model has an excellent recall score of 93.2%, meaning that it correctly recognised the vast majority of positive examples.

Adaboost Model: The accuracy of the Adaboost model was 93.5%. In terms of precision, the model achieved a score of 95.7% which indicates the proportion of true positive predictions among the positive predictions made. This model has a 91.8% recall, meaning that it correctly recognised 91.8 percent of positive examples.

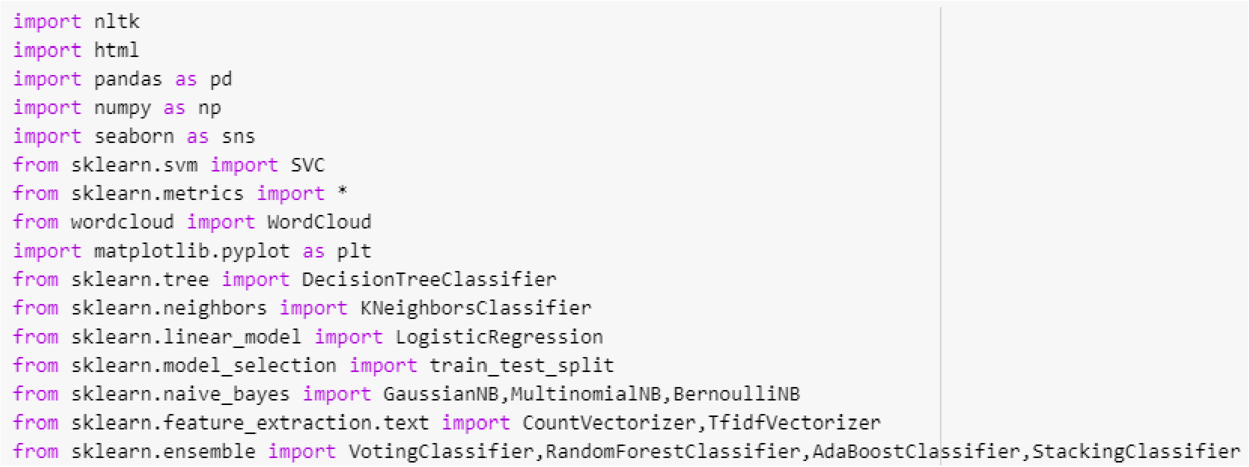
Hybrid Model: Finally, the hybrid model that was built using Logistic Regression, Random Forest and Adaboost models, was found to have the highest **accuracy of 97.85%.** This model has an accuracy score of 99.9%, which indicates that 99.9% of the time, the model generated accurate predictions. This model has the recall score of 95.9%, meaning that it correctly recognised 95.9% of positive examples.

Table 3 Results

Model	**Accuracy**	**Precision**	**Recall**
Logistic Regression	95%	99.9%	90.5%
Random Forest	96.4%	99.9%	93.2%
AdaBoost	93.5%	95.7%	91.8%
(Proposed)Hybrid Model	**97.8%**	**99.9%**	**95.9%**

The results of the evaluation of the models demonstrate the efficacy of the hybrid model in accurately classifying YouTube comments as spam or not spam. The high accuracy, precision and recall scores obtained by the hybrid model indicate its ability to effectively identify spam comments while minimizing the number of false positive predictions.

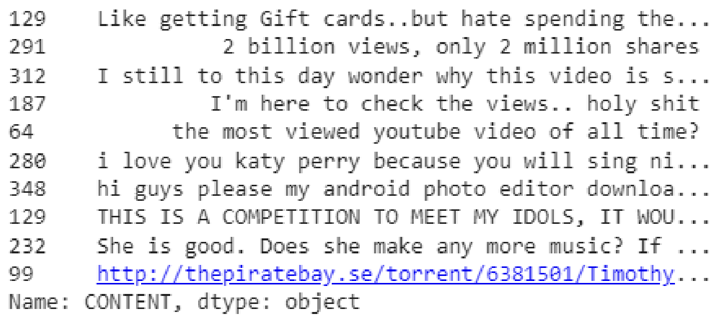
All packages shown in below figure which are used in this project.



```
import nltk
import html
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.svm import SVC
from sklearn.metrics import *
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB,MultinomialNB,BernoulliNB
from sklearn.feature_extraction.text import CountVectorizer,TfidfVectorizer
from sklearn.ensemble import VotingClassifier,RandomForestClassifier,AdaBoostClassifier,StackingClassifier
```

Figure 1 Packages used in the project

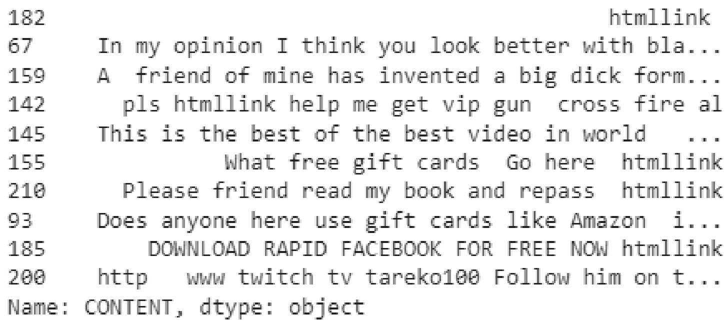
How data looks before preprocessing.



```
129    Like getting Gift cards..but hate spending the...
291              2 billion views, only 2 million shares
312    I still to this day wonder why this video is s...
187             I'm here to check the views.. holy shit
64           the most viewed youtube video of all time?
280    i love you katy perry because you will sing ni...
348    hi guys please my android photo editor downloa...
129    THIS IS A COMPETITION TO MEET MY IDOLS, IT WOU...
232    She is good. Does she make any more music? If ...
99     http://thepiratebay.se/torrent/6381501/Timothy...
Name: CONTENT, dtype: object
```

Figure 2 Data before cleaning

Data after cleaning & preprocessing.



```
182                                             htmllink
67     In my opinion I think you look better with bla...
159    A  friend of mine has invented a big dick form...
142      pls htmllink help me get vip gun  cross fire al
145    This is the best of the best video in world   ...
155              What free gift cards  Go here  htmllink
210      Please friend read my book and repass  htmllink
93     Does anyone here use gift cards like Amazon  i...
185        DOWNLOAD RAPID FACEBOOK FOR FREE NOW htmllink
200    http   www twitch tv tareko100 Follow him on t...
Name: CONTENT, dtype: object
```

Figure 3 Data after cleaning

Word Cloud

Most Used words in Spam Comments:

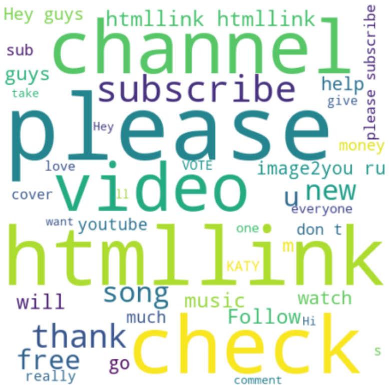


Figure 4 Most Used words in Spam comments

Most Used words in Normal Comments:

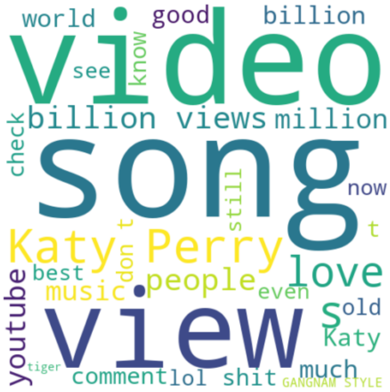


Figure 5 Most Used words in Normal comments

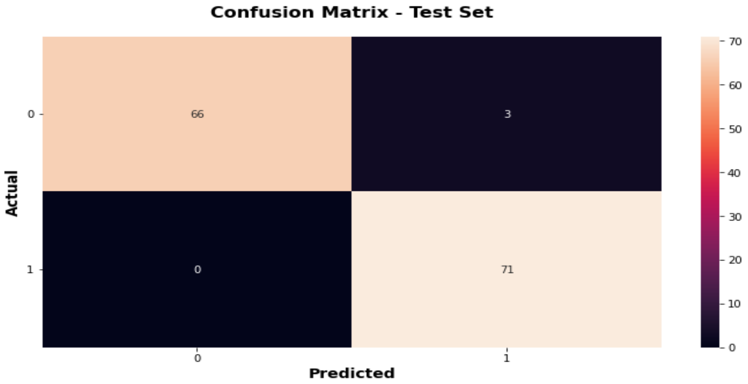


Figure 6 Confusion Matrix

In addition to these metrics, a confusion matrix was also used to evaluate the model. The effectiveness of the binary classifier can be summarised in a table by comparing predicted values to actual values; this table is called a confusion matrix. There were 66 correct predictions, 71 correct negative predictions, 0 incorrect negative predictions, and 3 incorrect positive predictions in the hybrid model's prediction matrix. The model accurately identified 66 comments as spam, as shown by the true positive predictions, and 71 comments since not spam, as indicated by true negative predictions. Incorrectly marking 3 comments as spam indicates a false positive forecast, whereas not identifying any comments as spam indicates a false negative prediction.

In conclusion, the confusion matrix provided a clear picture of the performance of the hybrid model and showed that it was able to accurately classify comments as spam or not spam. The high accuracy, precision, and recall scores indicate that the model was able to effectively identify spam comments and minimize false positive and false negative predictions.

Comparing the Proposed method with Existing Work

In comparison to an existing work, our best performing model, the hybrid model, demonstrates superior results with an accuracy of 97.8% compared to the logistic regression model in the existing work with 95.40% accuracy. This improvement in accuracy can be attributed to the combination of three base models, Logistic Regression, Random Forest, and Adaboost, using a voting classifier. The utilization of multiple models helps to tackle the problem from different angles and reduces the chances of overfitting and underfitting. The hybrid model's high precision and recall scores of 99.9% and 95.9% respectively, further solidify its superiority and effectiveness in detecting YouTube spam. These results highlight the significance and effectiveness of the proposed hybrid model in detecting YouTube spam compared to existing methods.

Table 4 Model Comparing with existing work

Existing Work	Accuracy Score
Logistic Regression [20]	95.40%
Proposed work	**Accuracy Score**
Random Forest + **AdaBoost +** **Logistic Regression** **Hybrid Model**	**97.85%**

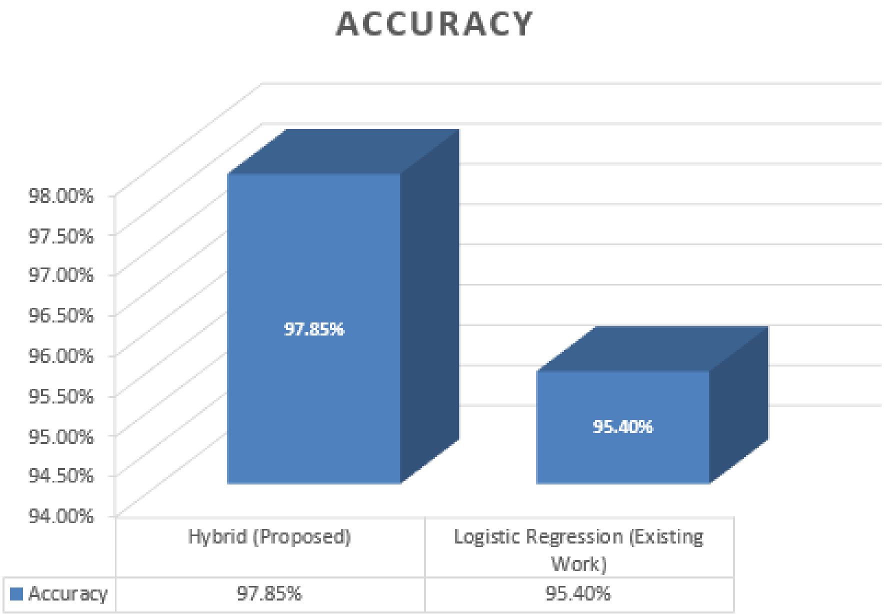


Figure 7 Accuracy of proposed model

As show in the figure Hybrid model has high accuracy than existing Logistic Regression.

5. Conclusion

In conclusion, the goal of this thesis was to develop and evaluate a machine learning model for detecting spam on YouTube. To achieve this goal, followed a structured process that involved data preprocessing, model development, data collection, and model evaluation.

In this Research, presented a study on the classification of the YouTube comments as spam or not spam. The study involved the preprocessing of a dataset that was collected from YouTube comments and the development of a machine learning-based model to classify these comments. The preprocessing involved basic data cleaning, removal of null and duplicate values, filtering of columns that were not relevant to the classification, removal of HTML tags and finally, converting the text into numbers using the CountVectorizer.

The study also involved the evaluation of several machine learning algorithms, including Decision Trees Classifier, Logistic Regression, Ada Boost Classifier, Random Forest, KNeighbors Classifier, Support Vector Machine, and Naive Bayes. However, the best performing model was found to be a hybrid model that combined Logistic Regression, Random Forest and Ada Boost Classifier using a voting classifier. The hybrid model achieved an accuracy of 97.8%, a precision of 99.9% and a recall of 95.9%.

The results of this study designate that the hybrid model is a reliable method for detecting YouTube comments that are spam. However, there are also some limitations of this study, including the use of a relatively small dataset and the absence of a comprehensive dataset that includes comments from different languages and cultures.

In conclusion, this study provides a valuable contribution to the field of spam detection and provides a basis for future work in this area. Future work should include the extension of this study to include a larger and more diverse dataset, as well as the evaluation of other machine learning algorithms and methods. Additionally, it would be valuable to extend this study to include the detection of spam in other forms of online communication, such as emails and social media posts.

6. References

[1] Sharmin, Sadia, and Zakia Zaman. "Spam detection in social media employing machine learning tool for text mining." 2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). IEEE, 2017.

[2] Alberto, Túlio C., Johannes V. Lochter, and Tiago A. Almeida. "Tubespam: Comment spam filtering on youtube." 2015 IEEE 14th international conference on machine learning and applications (ICMLA). IEEE, 2015.

[3] Alias, Nabilah, et al. "Video spam comment features selection using machine learning techniques." Indones. J. Electr. Eng. Comput. Sci 15.2 (2019): 1046-1053.

[4] Liu, Chen, and Genying Wang. "Analysis and detection of spam accounts in social networks." 2016 2nd IEEE International Conference on Computer and Communications (ICCC). IEEE, 2016.

[5] Trivedi, Shrawan Kumar. "A study of machine learning classifiers for spam detection." 2016 4th international symposium on computational and business intelligence (ISCBI). IEEE, 2016.

[6] Shirani-Mehr, Houshmand. "SMS spam detection using machine learning approach." unpublished) http://cs229. stanford. edu/proj2013/Shir aniMeh r-SMSSpamDetectionUsingMachineLearningApproach. pdf (2013).

[7] Sun, Nan, et al. "Near real-time twitter spam detection with machine learning techniques." International Journal of Computers and Applications 44.4 (2022): 338-348.

[8] Ahmed, Naeem, et al. "Machine learning techniques for spam detection in email and IoT platforms: analysis and research challenges." Security and Communication Networks 2022 (2022).

[9] GuangJun, Luo, et al. "Spam detection approach for secure mobile message communication using machine learning algorithms." Security and Communication Networks 2020 (2020).

[10] Kumar, Nikhil, and Sanket Sonowal. "Email spam detection using machine learning algorithms." 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2020

[11] Wu, Tingmin, et al. "Twitter spam detection based on deep learning." Proceedings of the australasian computer science week multiconference. 2017.

[12] Kontsewaya, Yuliya, Evgeniy Antonov, and Alexey Artamonov. "Evaluating the effectiveness of machine learning methods for spam detection." Procedia Computer Science 190 (2021): 479-486.