# MPSTME NMIMS

# Youtube Spam Comment Detection

## PROJECT SYNOPSIS

Final year IBM Project

## B TECH

CS

SUBMITTED BY

Pranav Kolhe B228

GUIDED BY

Prof. Dhananjay Joshi

# Title Page

Name of Student – Pranav Kolhe

Class Roll No. – B228

Present official Address with Email, telephone No. – pranav.kolhe035@nmims.edu.in, 8369612873

Branch - B Tech

Batch - 2019-2023

# 

# Content Page

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Topic** | **Page** |
| 1 | Introduction | 4 |
| 2 | Literature Survey | 5 |
| 3 | Methodology | 6 |
| 4 | Facilities required | 7 |
| 5 | References | 8 |

**Introduction**

YouTube, the world’s largest video sharing site, was founded in 2005 and acquired by Google in 2006. YouTube has grown tremendously as a video content platform, with the recent shift in online content to video.

At present, more than 400 hours of video are uploaded and 4.5 million videos are watched every minute on YouTube. It is easy for users to watch and upload videos without any restrictions.

This great accessibility has increased the number of personal media, and some of them have become online influencers.

YouTube creators can monetize if they have more than 1,000 subscribers and 4,000 hours of watch time for the last 12 months. Accordingly, spam comments are being created to promote their channels or videos in popular videos.

Some creators closed the comment function due to aggression such as political comments, abusive speech, or derogatory comments not related to their videos.

YouTube has its own spam filtering system, though there are still spam comments that are not being caught.

With the growth of Youtube, there is a growth of Spam Comments on it too.

Youtubers don’t have the means to address that, but it can be addressed through data science and coding.

Through this project, I will try to do the same and hope to improve on the previously implemented projects.

# Literature Survey

Here are three papers related to youtube Spam Comment detection, along with a brief summary of each and their contributions:

* "Youtube Spam Comments Detection" by Oviya Selvaraj, Anuradha Konatham, Dr. Paavai Anand. (2020)

In this paper, the spam comments detection methodology has been carried out using four Artificial Intelligence estimations – Logistic Regression, Ada Boost, Decision Tree and Random Forest. With the use of Neural Network, we can achieve an exactness of 91.65% and beat the present course of action by around 18%. The most notable AI procedures (Bayesian portrayal, k-NN, ANNs, SVMs) and of their suitability to the issue of spam.

* " An Approach for Spam Detection in YouTube Comments Based on Supervised Learning" by Amir Ali, Muhammad Zain Amin et al. (2016)

In this paper, they have evaluated several top-performance classification techniques for such purpose. The statistical analysis of results indicates that the Multilayer Perceptron and Support Vector Machine show good accuracy results of 94.89%.

* " A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model” by Hayoung Oh et al. (2021)

# This paper proposes a technique to detect spam comments on YouTube, which have recently seen tremendous growth. YouTube is running its own spam blocking system but continues to fail to block them properly. Therefore, we examined related studies on YouTube spam comment screening and conducted classification experiments with six different machine learning techniques (Decision tree, Logistic regression, Bernoulli Naïve Bayes, Random Forest, Support vector machine with linear kernel, Support vector machine with Gaussian kernel) and two ensemble models (Ensemble with hard voting, Ensemble with soft voting) combining these techniques in the comment data from popular music videos - Psy, Katy Perry, LMFAO, Eminem and Shakira

# Methodology

Depending on how it's set up, the way a youtube spam comment detection system works can be different, but here are some general steps:

* Problem Definition: Define the problem and the requirements for the youtube Spam Comment detection system. Identify the key features and performance metrics, such as accuracy, precision, and recall.
* Data Collection: Collect a dataset of comments on youtube videos with and without spam comments. The dataset should be diverse and representative of the target population.
* Data Preprocessing: Preprocess the dataset by choosing the relevant data to focus on the spam comments’ features.
* Model Selection: Select a suitable machine learning algorithm for the youtube Spam Comment detection task, such as Support Vector Machines (SVMs), Random Forests, or K-Nearest Neighbor (k-NN). Train and validate the model on the preprocessed dataset, using cross-validation techniques to evaluate its performance.
* Model Optimization: Optimize the model by tuning its hyperparameters, such as learning rate, regularization strength, and batch size. Use grid search or other optimization techniques to find the best combination of hyperparameters.
* Testing and Evaluation: Test and evaluate the youtube Spam Comment detection system on a new dataset of comments. Use performance metrics such as accuracy, precision, and recall to evaluate its performance. Iterate on the design and implementation of the system based on the results of testing and evaluation.

# Facilities required

### Hardware:

The hardware needed for a youtube spam comment detection system can vary depending on how it is used, but here are some general guidelines:

* **Processor:** A powerful processor is needed to process the data feed in real-time and perform the necessary machine learning algorithms. A quad-core processor with a clock speed of at least 2.5 GHz is recommended.
* **Memory:** Sufficient memory is required to store the data feed and perform machine learning algorithms. At least 8GB of RAM is recommended.
* **Storage:** Adequate storage is necessary to store the data feed and machine learning models. At least 500GB of hard drive space is recommended.
* **Graphics Card:** A dedicated graphics card is helpful for data processing and machine learning algorithms. A minimum of 2GB of dedicated RAM is recommended.

### Software:

Depending on how the system is put together, the software requirements for a drowsiness detection system can be different, but here are some general requirements:

* **Integrated Development Environment (IDE):** An IDE such as PyCharm, Eclipse, or Visual Studio Code is necessary for coding, debugging, and testing the software components.
* **Programming Language:** A programming language such as Python, C++, or Java is required for implementing the software components.
* **Machine Learning Libraries:** Machine learning libraries such as TensorFlow, Keras, or PyTorch are required for implementing the machine learning algorithms.

# References

* <https://www.irjet.net/archives/V7/i4/IRJET-V7I488.pdf>
* H. Oh, "A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model," in IEEE Access, vol. 9, pp. 144121-144128, 2021, doi: 10.1109/ACCESS.2021.3121508.
* Ali, Amir & Amin, Muhammad. (2016). An Approach for Spam Detection in YouTube Comments Based on Supervised Learning.