International Journal of Gender, Science and Technology

..ISSN: 2040-0748

UGC Care Group I Journal

Vol-12 Issue-02 Aug 2023

Applying Machine Learning Techniques and Detecting Spam in YouTube Comments

[1]Dr T. BENARJI, [2]P. RENUKA

[1]Professor, Department of CSE, Indur Institute of Engineering and Technology, Siddipet, Telangana, Hyderabad. Email: tharinibenarji@gmail.com

[2]Associate Professor, Department of ECE, Indur Institute of Engineering and Technology, Siddipet, Telangana, Hyderabad. Email: renoostar@gmail.com

Abstract: *Spammers have taken advantage of the growing sophistication of social media websites by flooding video comment sections with spam in an effort to trick viewers into engaging in harmful actions. This job involves collecting comments from YouTube and detecting spam. Tools like YouTube Bookmaker and Google Safe Browsing may identify and filter spam videos on YouTube. These solutions will prevent harmful links from being accessed, but they won't be able to safeguard the user in real-time. Researchers and businesses have therefore taken radically divergent tacks in their pursuit of a social media platform devoid of spam. Using four AI estimations—Logistic Regression, Ada Boost, Decision Tree, and Random Forest—a poll was conducted to choose the approach for spam comment identification. We can outperform the current procedure by around 18% and get an accuracy of 91.65% with the help of Neural Network. The most prominent artificial intelligence methods and how well they work on the spam problem: Bayesian representation, k-NN, ANNs, and SVMs.*

I. INTRODUCTION

In recent years, individuals' day-to-day lives have been increasingly enriched by informal online communities such as Facebook and YouTube. To keep in contact with loved ones and to publish one's own ideas and opinions in online journals, people use social media as a kind of online community. As a result of this trend, these platforms attract a flood of users, making them prime targets for spammers. YouTube has surpassed all other youth-oriented informal communities in terms of popularity. Bloggers who call themselves "beauty gurus" or "beauty influencers" have created a plethora of cosmetic tutorials, the vast majority of which cater to adolescent females. Present day YouTube content (videos) production numbers 400 million each day, with 200 million customers. Spammers may take use of YouTube's vast ecosystem to target consumers with unrelated material. The goal of these spam my messages is to trick people into visiting dangerous websites that contain viruses, phishing, and other forms of online fraud. Among YouTube's most prominent features is the comments area that appears under each user-uploaded video. With this function, people may voice their thoughts and views.

This research employs the idea of machine learning, which is a subset of artificial intelligence, to forecast the presence of spam remarks in the comments area of YouTube videos. A massive amount of tagged datasets is essential to the supervised learning method. In order to forecast the spam comment, the suggested classification technique (Logistic Regression) is used. Project goals include providing a high-level overview of machine learning methods and outlining a method for making predictions. Machine learning offers a new way to investigate and improve the accuracy of predictions, and it's far better than traditional data analysis methods.

Typically generated by automated bots posing as human clients, spam comments are often irrelevant to the provided video. Spammers sometimes target the comments area to offer unrelated remarks, links, and ideas. The goal of artificial intelligence (AI) is to sift through massive amounts of data, identify relevant instances, and then transform that data into a valid structure for further use by predicting its future use. There are two ways to break down data: grouping and expectation. Grouping shows main types of data, while anticipation predicts patterns in future data. The constructive view of the videos' substance will be ruined by the nasty spam comments. The plan to foresee spam comments has begun, however it has

not been fully developed or finalized to provide an accurate prediction of spam comments.

II. RELATED WORKS

2.1 An effective technique for detecting communities in social networks based on modularity, published in the IEEE.

In a social network (SN), the goal of the network identification procedure is to find clusters of nodes that are more closely linked to one another than nodes that are outside the cluster. One of the challenging difficulties in the era of big data analysis, particularly in the field of long-distance interpersonal contact, is this technique. The usage of hubs to represent on-screen actors and edges to represent relationships among the entertainers is commonplace when communicating with SN using diagram information structure. While SNs do include certain computations for network identification purposes, these methods all have their limitations when used to networks with extensive coverage. In this paper, we provide a competent measured quality based network discovery algorithm. By comparing it to other network identification computations that use the most well-known informal organization datasets, the suggested method has been thoroughly tested. Various metrics such as particularity, bunching coefficient, execution duration, and others have been used to examine the calculation's execution.

2.2 Distributed Louvain Algorithm: A Scalable Method for Detecting Large-Scale Graph Communities.

Another popular network discovery computation for huge diagrams based on the Louvain approach is shown here. To keep everyone on the same page and make sure that processors are changing their communication, we employ a communicated delegate allocation. In addition, we devise an additional heuristic method to purposefully aid the network architecture in an appropriated area, and to ensure the union of the circulating bunching computation. The flexibility and accuracy of our computation using designed chart datasets have been shown in our escalating test study.

2.3 Keeping "Which Videos Are Similar to You? Improving Video Recommendation via Mastering a Common Attributed Representation.

However, learning such a typical depiction in insufficient client video communications, managing the frosty starting problem, and adjusting the tasks of social qualities and substance characteristics are still challenging. An approach based on lattice factorization known as regularized dual-factor regression (REDAR) should be proposed immediately. There is now a sufficient abuse of social and substance data to alleviate the sparsely problem, and there is a dexterous joining of characteristics and substance attributes. The goal of developing a more gradual version of REDAR is to address the problem of cool beginnings. We test the proposed method extensively on a real-world interpersonal organization dataset for video recommendation applications, and the results demonstrate that, on average, the method outperforms state-of-the-art pattern techniques by more than 20%.

22.4"Search algorithms for scalable communities using parallel heuristics."

Locating networks has become a crucial task in many diagrammatic theoretical contexts. Its purpose is to reveal "normal divisions that exist" within real systems, without imposing size or cardinality constraints on network topologies. The core algorithms are notoriously unexpected and inherently sequential, which limits their use to network identification for very wide-area PCs, despite the technology's promising future. Now, heuristics for parallelization that use the Louvain technique as a sequential layout are available for fast network location. One such iterative heuristic for "measured quality enhancement" is the multi-stage Louvain approach. The method's ability to quickly and efficiently discover high particularity network segments has contributed to its rising profile since its creation by Blondel et al. (2008). Though it seems different compared to the previous louvain usage, our same execution may provide organized returns of better approximated quality for the vast majority of data.

III. METHODOLOGY

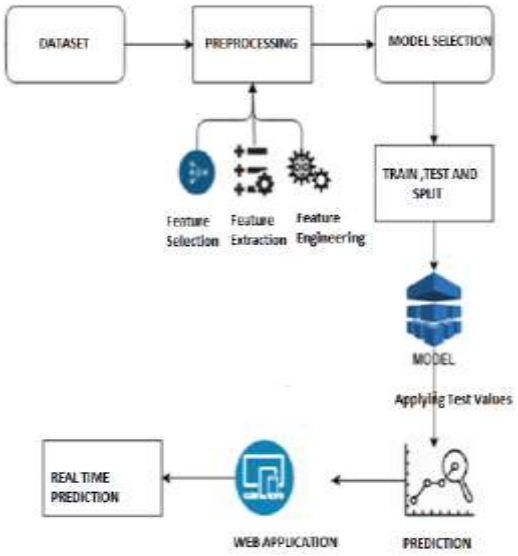


Fig 1. System Diagram

3.1 Dataset: Incorporating these terms from the characteristic-set based on their entropy score has allowed us to reduce uncertainty in the prediction findings. This is because these phrases significantly impact the frequency count in both spam and non-spam YouTube.

Preprocessing the communications is an essential first step before beginning preparation (3.2). The first rule is that everything must be lowercase. It is incorrect to treat as two distinct words a term that appears in both uppercase and lowercase. After that, every message in the dataset has to be tokenized.

3.3.3 Feature Selection: Using the terms in the dataset may reduce uncertainty in the prediction of the final results, which is great since such phrases have a surprising influence on the frequency count in spam and ham comments on YouTube.

3.4 Feature Engineering and Extraction: One supervised characteristic that ranks attributes in a step-by-step way based on their relevance in forecasting a goal is attribute significance. This is where Count Vectorizer comes in handy; it takes a "set of text documents and turns them into a matrix of token counts. The following method is applied to this:

3.4.1 N-grams: N-grams are used to enhance precision. This only applies to singular words; however, the whole meaning changes when two mutual terms are present. Splitting text into tokens of two or more words, as opposed to being a single word, results in better variety in accuracy.

If the feature should consist of n-grams of words or characters, the analyzer will determine it in 3.4.2. When using the 'char_wb' option, character n-grams are only generated from text that is within word borders. Any n-grams that are outside of words are simply filled with space.

3.4.3 Words: "Either an iterable over words or a mapping (like dicts) where terms are keys and values are indices in the feature matrix... When not explicitly stated, the input texts are used to establish the vocabulary. The mapping's indices must be unique and uninterrupted from zero to its maximum value.

3.4.4 Model Construction

Following Preprocessing, a method must be devised for version construction that preserves the project function's abilities in line with the labeled model, which is constructed according to the Supervised set of rules.

If "max features" is not None, then construct a vocabulary using the top max features in order of word frequency in the corpus. If vocabulary is not None, then this option will be disregarded.

When it comes to addressing problems, the boosting method that has been used is Adaboost.Bringing together many poor classifiers into one strong one helps. Decision stumps, short for "decision tree with one split," are the poor learners from whom it first separates. After that, it sorts the datasets according to difficulty, giving more weight to the cases that were more challenging and less weight to the ones that were handled well. After dividing the decision stumps into two groups, we'll choose a threshold value; all of the data will fall into one of these categories. Due to its failure when presented with a result that deviates significantly from the threshold, it achieves only mediocre accuracy on the dataset.

With the use of a decision tree, we may ask a succession of yes/no questions about our data, which will lead to a continuous value or prediction. This is when it seeks to construct nodes that include a large percentage of data points from a specific or single class by determining the values of characteristics that partition the data into segments. Because of over fitting in the data, this nonlinear model isn't as accurate as other algorithms, but it's designed with

numerous linear limits; here, we offer a model labels and features so it can grasp how to categorize points depending on attributes.

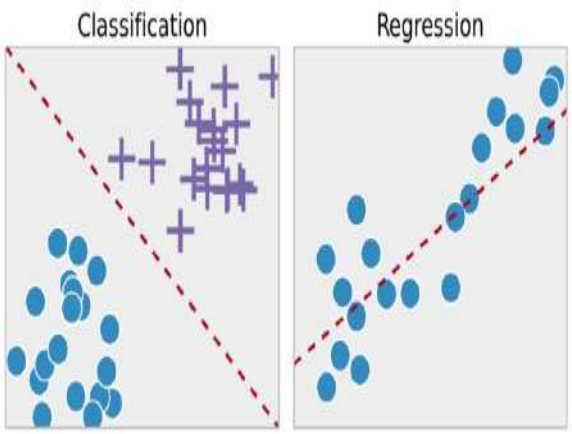
Random forest is not as accurate as other algorithms since it combines several blocks of decision trees into one.

Predicting a variable's binomial or multinomial value is the job of logistic regression. To determine the result, it employs a statistical method. The results are of a binary kind. The output here will be correct either x or y since it utilizes a logit function to forecast the chance of occurrence of binary outcomes, which follow the bernoulli distribution. In this case, it analyses the dataset and draws the conclusion that x or y is spam or ham.

IV. TECHNICAL MODULE

Following preprocessing, the project's architecture should preserve the function's capacity to generate a label model. A Supervised set of rules may be used to build this model.

Supervised learning:



A learning function, it uses tagged training data to infer an output and a mapping between the two. We need to find the best model parameters to search unknown labels on different devices (test set) by fitting them into the classified training set. We refer to it as the venture regression if the label is a real variety.

AdaBoost

The term "AdaBoost," which stands for "adaptive boosting," describes an algorithm that works well when combined with others. For the purpose of improving the model's prediction ability, it is a method that chooses which characteristics to enhance. It speeds up the model's execution by removing superfluous features.

Decision Tree Model

This approach solves the issue by modeling it in a tree structure, where each node represents an attribute and each leaf node specifies a class label.

Unsupervised Learning

It is a paradigm for ensemble learning that makes use of many decision trees in tandem. It is composed of individual pieces that are decision trees.

Regression using Logistic

A logistic function known as logit is used by this technique, which employs a statistical approach to determine a binary result. If you want to model many types of events, such whether to find a picture of its characteristics, you can.

ALGORITHMS	ACCURACY
Logistic Regression	0.9540
Decision Tree	0.5438
Random Forest	0.8469
Adaboost	0.7125

Figure 3: Graphing Performance and Analysis

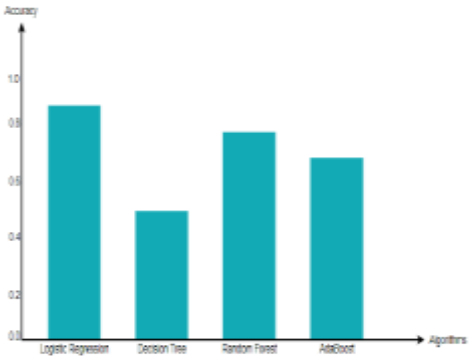


Fig 4. Analysis graph

V. RESULTS AND SCREENSHOTS

The findings show that there is a large pool of demonstrated representation tactics for identifying and removing spam from YouTube comments. Quite a few of them really have the option to attain accuracy rates over 90% with no or no obstructed ham whatsoever.

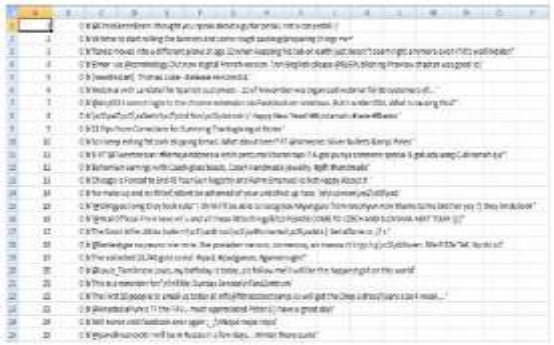


Fig 5. Dataset

Fig 6. Home Page

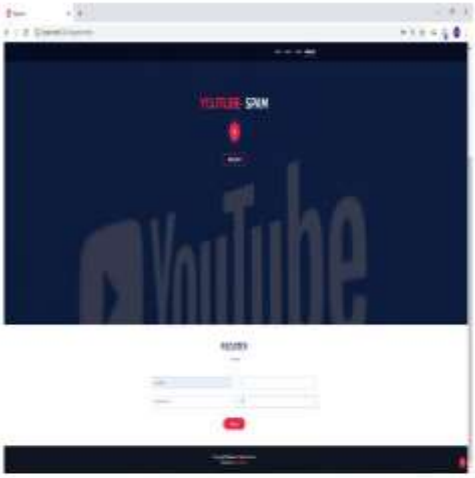


Fig 7. Registration Page of the Project

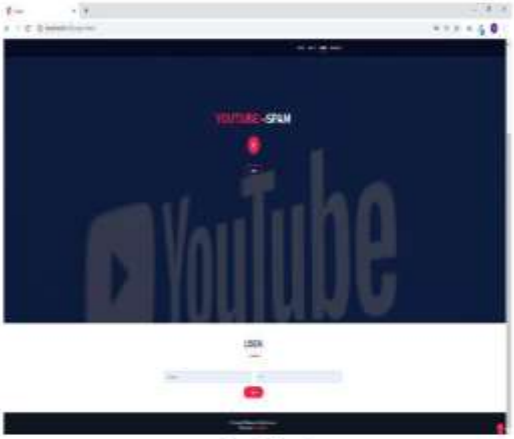


Fig 8. Login

Fig 9. Comments Section of the Video

Fig 10. Result Page

CONCLUSION

Several methods are used to categorize YouTube comments as spam or not spam (ham). The results obtained from testing this method using real-time comments on YouTube show that it is 18% more accurate than the current method. Since the YouTube API is accessible to everyone, it has the potential to influence spammers' actions over time. The reality is that YouTube's spam function is always evolving at a dizzying rate.

REFERENCES

[1] P. Chopade, J. Zhan, and M. Bikdash. Node attributes and edge structure for large-scale big data network analytics and community detection. In International Symposium on Technologies for Homeland Security (HST), pages 1–8, 2015.

[2] X. Que, F. Checconi, F. Petrini, and J. A. Gunnels. Scalable community detection with the louvain algorithm. In Parallel and Distributed Processing Symposium (IPDPS), pages 28–37, 2015.

[3] P. Cui, Z. Wang, and Z. Su. What videos are similar with you?:Learning a common attributed representation for video recommendation. In ACM International Conference on Multimedia (MM),pages 597–606, 2014.

[4] H. Lu, M. Halappanavar, A. Kalyanaraman, and S. Choudhury. Parallel heuristics for scalable community detection. In International Parallel & Distributed Processing Symposium Workshops (IPDPSW), pages 1374–1385, 2014.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] S. Oreg and N. Sverdlik. Source personality and persuasiveness:Big five predispositions to being persuasive and the role of message involvement. Journal of Personality, 82(3):250–264, 2014.