A COMPARISON OF MACHINE LEARNING ALGORITHMS FOR INDIAN YOUTUBE SPAM FILTER

¹Hasan Asif and ²Masood Anzar

Department of Mechanical Engineering, Meerut Institute of Engineering and Technology

²Automation Engineer Paragon Aluminum Ltd., India

Email: { mailsforhasan@gmail.com }

Abstract: Expressing your view on social media has become part of our daily life and one of the most used media is the YouTube comment section, where the user can critic any video published by the creator. With the incessant augmentation in the users of YouTube, the quantity of spam comment has also rose meteorically. In this paper, we propose and developed a comment spam filter, it was achieved by using a supervised learning approach which was trained on our self-generated dataset. In addition to this, literature review of different method used by researcher to identify the spam content is also presented. Furthermore, paper illustrates the description and evaluation of accuracy, precision, and recall of different Machine Learning Technique performance such as Linear classifier, Random forest classifier, AdaBoost classifier, Decision Tree, Naive Bayes, and SVM.

Keywords: Machine learning, NLP, Spam Filter, YouTube

1. Introduction

With the introduction of fibre optics and low rate in the telecom industry lead to the meteoric rise of traffic in YouTube India[1], These days, YouTube's traffic on mobile has increased dramatically to 85%, with major contribution on watch time (around 60%) on videos coming from outside of the six largest metros city in India.[2]. YouTube has emerged as a powerful platform for users, content creators and advertisers. A huge variety of content combined with the growing user base and popularity of the creator in the society— entice spammer on YouTube than in any other platform. It provides a good opportunity to a spammer to promote their channel/video/product. YouTube claims that during the period of July-September 2018 they have almost deleted 224 million malicious comments out of billion of comments published in this 3 months period[3].

Nowadays YouTube comes out to be a typical medium of entertainment, where users can scroll down to billion of videos. Whereas, creators get money for every video if they get monetize. Now it gets important for the creators to know the response of the viewer on that particular video, which can be traced from YouTube comment section. Due to the spam comments, many times creator missed the valuable comment. This paper, evaluated different machine learning algorithms for spam detection. To do so we chose the Supervised learning approach so, we have generated our own custom dataset containing both spam and non-spam comments. All Supervised learning algorithm used in this project are Logistic Regression, Naive Bayes, SVM, Decision Tree classifier, Random forest classification, and AdaBoost classifier. In the study, we have found that spam comments are of two types which are as follows

- 1:spam comment that is promoting product/video/channel.
- 2:writing comment irrespective of the video, these comment can contain malicious and curse word or things which are irrelevant to the video. Several research on spam classification has been done, some achieved high accuracy, while some are still going to due to its dynamic nature. Developing a spam classification mechanism is a demanding task, as spammer are actively changing their spamming

method [4].R.M.A Mohamed[4] discussed the

same issue that spammers are aware of the fact that, attacking the user with the same set of features, more probably the anti-spam parties might develop more robust tools for combating. In his study, an enhanced model is proposed for ensuring a lifelong spam classification model using the Enron-Spam dataset which contains 33,716 email messages, of which 16,545 email messages are marked as ham and 17,171 email messages are marked as spam. In 2011 [5] paper reviewed some of the most popular machine learning methods. Illustration of various algorithms, in addition to this, performance of different algorithms juxtapose against each other, on 6000 emails dataset which contains spam and non-spam mails. In term of accuracy, Naive Bayes algorithm comes out to be the best as compared to the other machine learning and shallow deep learning models. one of the recent paper [6] on spam classification presented an evaluation of spam comments identification on various algorithms such as Ensemble model, Artificial Neural Nets, Trees and Regression. Paper discussed the applicability of above method in spam classifier. Achieve an exactness of 91.65%, With the use of a Neural Network. Saumya Goyal et al. [7] developed spam message detection using real Twitter social media dataset that is applied using KNN and decision tree. The metrics used for classification are precision, recall, F measure and class, FP rate, and TP rate. Simran Kanodia et al. [8] suggested a Markov Decision process for YouTube spam message detection and the result is compared to other data mining tools used in this field. The Markov Decision accuracy is 78.82 % which is quite better than those of other data mining algorithms out of which the maximum accuracy is obtained through the random forest which is 72.52 %.

Shreya Aiyar et al[9] , in his work, deployed peculiar method to identify the spam content in the comments. Typical machine learning algorithm were tested on N-gram approach, and results from various models(SVMs, NB, and RF) were compared with the baseline model(Naive Bayes) owing to its

simplicity and high classification accuracy[12]. Results demonstrated the efficacy of unique approach in spam identification. Several studies have proved the feasibility of Random Forest Classifier and Support Vector Machine in spam classification Task[13] and poses the ability to being employed on large dimension data, especially SVMs. Igor Santos et al. [10] while discussing spam as a threat to computer security in his paper, incorporated an anomaly detection approach to spam filtering. With this concept, author reduces the necessity of labeling spam messages and better accuracy was achieved. This approach transforms spam e-mails into word frequency vectors. Thereby, classifying an e-mail as spam or non-spam by calculating the distance from the base non-spam email.

M. McCord et al. [11] In his research, various classification models were compared for the spam detection task. Random forest classifier was discovered as the most efficient in that particular task. They trained and tested their algorithms with Twitter data with content and user-centered facets.

2. Corpus and Data Preprocessing

Data is generated over a period of 10 days using YouTube API. We randomly selected a popular video from 17 different YouTube channels in India. A total of 1700 comments were collected. Fig.1 shows the histogram of comments word count, which is clearly right skewed as expected. This represents that very few people write comment more than 200 words.

Fig. 1: Histogram

Fig. 2: Class vs length of comment

It is clear from Fig. 2 that there is a huge discrepancy in the mean length of spam and non-spam comment.

The average length of spam comment comes out to be 79 words whereas, for non-spam comment it is just below 43 words. Collected comments were very messy and multilingual. So the first stage required cleaning the data, in which the removal of comments in other languages except English is done. After the cleaning step 500

comments were left, the next step was to clean those comments to do so, we applied Text cleaning using NLTK library which includes

- ✧ Removing everything except alphanumeric letter and some special symbol like[: , /]
- ✧ Tokenization
- ✧ Removing Stopwords
- ✧ Converting all letter into lower case.

The Summary of the dataset is provided in table 1 in detail. After cleaning the whole dataset, each comment was to be converted into a number, for our work we selected Term Frequency Inverse Document Frequency (TF-IDF), it is a unique way to transform the text documents into numbers, to make it comprehensive to computer, as computer can understand numbers only. By leveraging TF-IDF in Natural Language Processing each comment in the corpus was represented by the row in matrix and every word in the comment represented by the columns. During the model training process the count of column (Vocabulary size) was defined and which can be altered as per requirement.

During the text preprocessing stage, application of TF-IDF method gives an output in the form of matrix by storing the frequently occurring words as the column and each row was filled by the number of times each word repeated in the sentence. The benefit of using TF-IDF as compared to the Bag of Word method that it penalized those word which occurs frequently in the document.

Table 1 Summary of the dataset

Total number of comments before cleaning	1700
Total number of comments after cleaning	500
Mean length of comments	61 words
Mean length of non-spam comments	43.94 words
Mean length of spam comments	79.76 words
Number of distinct user	473
Number of spam comments	235
Number of non-spam comments	265

3. Performance Metrics

To assess the performance of different algorithms we calculated Precision, Recall, and Classification Accuracy score. One thing should be kept in mind in the case of spam classification, which is to keep False Positive (FP) as low as possible. Why is this so? Suppose a non-spam comment is passed through the model but it is erroneously classified as spam. Due to which useful information might be lost. On the other hand, we can be a little less harsh to False Negative (FN).

A true positive (TP) in confusion matrix represents the number of times model predicts comment as spam, and in actual it is spam too. Likewise, a true negative (TN) is represents the number of times model predicts comment as non-spam, and in actual it is non-spam too.

A false negative (FN) shows the number of times model predicts comment as non-spam, and in actual it is spam. A false positive (FP) show the number of times model predicts comment as spam, and in actual it is non-spam [14].

Different models are contrasted based on the following metrics which are discussed in detail below [15,16]

Classification Accuracy: Accuracy of the model vaguely define how well the model is performing in identification of each class. In mathematical terms, classification is defined as the sum of TP and TN over the sum of TP, TN, FP, an FN. It is preferred to have classification as close to 1.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision: It is describe as the ratio of TP and the sum of TP and FP. It is implied as the measure of the correctly identified positive cases from all the predicted positive cases. In more clear words how many of predicted spam by model is correct. For spam filtering, precision should be as high as possible. To achieve this we need to reduce FP as much as possible or in other word the best model would be one which has less FP because we don't want to reduce any valuable information.

$$Precision = \frac{TP}{TP + FP}$$

Recall: it is the ratio between TP and the sum of TP and FN. It is the measure of the correctly identified positive cases from all the actual positive cases.

$$Recall = \frac{TP}{TP + FN}$$

4. Description of Classifiers

Logistic regression

It is an algorithm that uses a linear approach to find the prediction class that is binary in nature it uses a logistic function $h(x) = 1 / (1 + e^x)$ This forms a S-shaped curve.

Logistic regression predictions are discrete values (i.e., 1/0) after applying a transformation function. Due to the "logistic" function used in transformation it is called as Logistic regression algorithm [18]

Despite having good accuracy in spam classification task, logistic regression poses some demerit. Logistic regression gives equal importance to every feature during the model

training process. Hence, as a result model cannot reach to the global minima and cannot deliver the best output in identification of the classes during the testing stage. For Linear Regression cost function calculated by Mean Squared

Error (MSE), while Logistic Regression leverage Sigmoid function. Application of Logistic Regression can be found in diverse field for classification task due to its simplicity and demonstrated good accuracy in spam identification [17].

Naive bayes

Naive Bayes (NB) is a supervised machine learning algorithm that bases on Bayes law In other word, if you know the probability of certain event but do not know the probability of other events, so in order to calculate the probability of other event Bayes law use the probability of those certain event. It is called naive because it assumes that each feature is independent but in actual these features might be mutually dependent. This makes the algorithm simple yet effective. Naive Bayes classifier is used in news segmentation, image classification, email categorization and in many tasks. Training of data is quite fast in Naive Bayes because it calculate only probability of different class with combination of inputs is calculated.

As shown in equation below which represents Bayes theorem mathematically, In classification problems, we use it to calculate the classification output(y) given some input feature(x).

$$P(x | y) = \frac{P(y | x) * P(x)}{P(y)}$$

SVM:

SVM is one the most effective classifier among the linear category. The mathematical intuition behind SVM is to maximize the margin width among all decision surface and this denoted by hyperplane in Fig 3.

Distance of closest data point to the decision surface is called margin. And, the points closest to the decision surface are known as support vectors, there should be minimum 2 of these support vector present in SVM classifier. SVM prevents overfitting and model is computationally less attractive even on large datasets [20,21]. Verma et al proposed a system for spam classification using Support Vector Machine algorithm

Fig. 3: Random forest

Random forest its effectiveness can be witness in discreet or continous output prediction. It uses decision tree underneath and forms multiple trees. In Random forest data set is divided into batch of random data sample further train

decision trees formed for each division. Takes majority vote out from individual decision tree, while with regarding to regression problem output is measured by calculating the mean of different trees output.[22]. DeBarr[25] chosen random forest algorithm for mail spam detection on RFC 822 dataset. During preprocessing stage, TF-IDF based feature extraction used. For robust classification model coupled with active learning. Despite, overfitting issue random forest are quite popular among non-linear classification model.

Decision tree is a popular non-linear algorithm, based on tree. Decision trees are pretty intuitive to work with, start at the top and work way down until we get to a point where we cannot go any further, and we reach to the classification output. The very top of the tree is called the root node, and following node are called intermediate note, lastly where no further division possible that node is called leaf node(predicted classification)[23]. To determining which feature is best for root node, we need to look into the impurity of each feature. There are a bunch of methods to calculate impurity, but the most possible and simple is "Gini". Feature with lowest gini impurity value chosen as the root node as it classify the data best among other features[24].

Adaboost algorithm leverage the boosting technique for classification and regression problems. Adaboost or boosting is a sequential process where in boosting technique different weak models are trained and minimize the error. Further, combined to create one robust classifier. To begin with, particularly in adaboost algorithm, each sample in training point has assigned the same weight equals to the reciprocal of total number of sample present in dataset. Then the next step would be to pick the classifier or the decision

MODEL	PARAMETER	VALUE
RANDOM FOREST	Maximum depth	80
	Maximum feature	sqrt
	Minimum sample leaf	2
	Minimum sample split	10
	n estimator	5
Decision Tree	Criterion	gini
	Minimum sample leaf	2
SVM	Kernel	linear

Table 2 Different model's Parameters

stump which gives the lowest error rate. Now, if the voting power for the classifier evaluated comes around the desirable range, we update the weight of each sample based on if it was classified correctly (weight goes down) or it was classified incorrectly(weight goes up and this loop iterate over till best possible adaboost classifier is achieved. Adaboosting technique is quite effective on large dataset and it took a computational tool while training.

5. Results

Despite having such a small dataset different models performed quite well in terms of accuracy. Chart 3. shows the accuracy of different models, in which highest accuracy was achieved by the SVM classifier, with the classification accuracy of 83% owing to its simple nature and less overfitting. whereas, Random Forest classifier achieved only

68% and performed quite unsatisfactory on detection of spam comments correctly. AdaBoost, Naive-Bayes, and Linear classifiers have more or less than 80% accuracy on the same dataset.

Fig. 4: Chart 3 Accuracy of different models

Fig. 5: Chart 4 Precision vs recall

Chart 4 illustrates that in terms of precision, Random Forest classifier, Linear classifier, and Naive-Bayes classifier reached around 95% respectively. whereas, recall for these model is quite low which means, these models fail in

correctly identifying the spam class and having recall value 44%,66%,66% for Random Forest classifier, Linear classifier, and Naive-Bayes classifier respectively. Precision of SVM is around 86% and recall is 84% which indicates less loss of information in identifying correct spam class. The Decision tree and AdaBoost classifier both have precision and recall more and less than 80%.

6. Conclusions

We have evaluated number of classical machine learning algorithms in this article and their applicability, to address the issue of Spam comment in YouTube video. Our classifier classifies comments into two groups spam and non-spam. we generated our own YouTube comment dataset using YouTube API and labeled all comments in the spam or non-spam category. Descriptions of the different algorithms are

presented, in addition to the performance metrics such as Accuracy, Precision, and

Recall were calculated on the dataset to find the best model. The algorithms showed very promising results, among all the six algorithms best result was achieved by Support Vector Machine. Some algorithm have effective outcomes and some are lacking in effectively predicting the correct output. Overall, this research study extends the support on supervised learning based spam classification system. It can be the appropriate approach for creating a spam filtering system. But still, the spam filtering system has some lacking, as spammer devises of a new way to permeate the filter which is the major concern.

References

- [1] Snigdha Poonam, How India conquered YouTube <https://www.ft.com/content/c0b08a8e-4527-11e9-b168-96a37d002cd3>, 2019
- [2] India-is-youtubes-largest-and-fastest-growing-audience-in-the-world-ceo
- [3] economictimes.indiatimes.com/industry/media/entertainment/india-is-youtubes-largest-and-fastest-growing-audience-in-the-world-ceo
- [4] YouTube: How many comments were removed in three months <https://www.bbc.com/news/newsbeat-46559772>, 2018
- [5] R. M. A. Mohammad, A lifelong spam emails classification model, Applied Computing and Informatics, <https://doi.org/10.1016/j.aci.2020.01.002>
- [6] W.A. Awad and S.M. ELseuofi, Machine Learning method for spam spam E-mails Classification, IJCSIT, Vol 3, No 1, Feb 2011
- [7] Oviya Selvaraj, Anuradha Konatham, Dr.Paavai Anand, YouTube spam comments detection, IRJET, Vol. 7, No. 4, Apr 2020
- [8] Saumya Goyal, Prof. R. K. Chauhan, Shabnam Parveen, Spam detection using KNN and decision tree mechanism in social network, presented at International conference of Parallel, Distributed and Grid Computing, 2016
- [9] Simran Kanodia, Rachna Sasheendran, and Vinod Pathari, A Novel Approach for Youtube Video Spam Detection using Markov Decision Process, DOI: 10.1109/ICACCI.2018.8554405
- [10] ShreyasAiyar and Nisha P shetty, N-Gram Assisted Youtube Spam Comment Detection in Proc, International Conference on Computational Intelligence and Data Science, 2018, pp.174–182
- [11] Santos, C. Laorden, X. Ugarte-Pedrero, B. Sanz, P. G. Bringas, Spam Filtering through Anomaly
- [12] Detection, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 203–216. doi:10.1007/978-3-642-35755-8 15
- [13] M. McCord, M. Chuah, Spam Detection on Twitter Using Traditional Classifiers, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 175–186. doi:10.1007/978-3-642-23496-5 13
- [14] H. Zhang, The Optimality of Naive Bayes, Proceeding of 7th International Florida Artificial Intelligence Society Conference, January, 2004.
- [15] M. Fernandez-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems?, Journal of Machine Learning Research, Vol.15, 2014, pp. 3133–3181.
- [16] Classification: Precision and Recall <https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall>, 2020.
- [17] R.M. Aliguliyev, R.M. Aliguliyev, S. Nazirova, Classification of textual e-mail spam using data mining techniques, Application of Computer Intelligence and Soft Computing Vol.8, 2011.
- [18] G.V. Cormack, M.D. Smucker, Efficient and effective spam filtering and reranking for large web datasets, Information. Retrieval, Vol.14, 2011, pp. 441–465.
- [19] Y. Han, M. Yang, H. Qi, X. He, S. Li, The improved logistic regression models for spam filtering, in: 2009 International Conference on Asian Language Processing, 2009, pp. 314–317, DOI: 10.1109/IALP.2009.74
- [20] Oviya Selvaraj, Anuradha Konatham, Dr.Paavai Anand, YouTube spam comments detection, IRJET, Vol.7, Apr 2020
- [21] Rennie JD, Shih L, Teevan J, et al. Tackling the poor assumptions of naive bayes text classifiers. In Proc of the 20th International Conference on Machine Learning, 2003, pp.616- 623.
- [22] S.O. Olatunji, Improved email spam detection model based on support vector machines, Neural Computing Application, 2017, pp.1–9.
- [23] Support vector machine, https://en.wikipedia.org/wiki/Support_vector_machine. Random forests classifier <https://www.datacamp.com/community/tutorials/random-forests-classifier-python>
- [24] Kajaare Das, Ravi Narayan B. , A Survey of Machine Learning: Concept, Algorithms and Applications, IJIRCCE, Vol. 5, No. 2, Feb 2017.
- [25] Rashid Chowdury, Md. Nuruddin Monsur Adnan, G.A.N. Mahmud, Rashedur M Rahman, A Data Mining Based Spam Detection System for YouTube, 2014 DOI: 10.1109/ICDIM.2013.6694038
- [26] DeBarr, D., & Wechsler, H., Spam detection using clustering, random forests, and active learning. In Sixth Conference on Email and Anti-Spam, 2009, pp. 1-6.