Predicting Human Well-Being Using Social Media Via Machine Learning

SHIKHA KAUSHIK[1], PRACHI SHARMA[2]
[1,2,3] *Department of Computer Science & Eng., Poornima College of Engineering, Jaipur, Rajasthan, India*

Abstract -- Web-based social networking can be named as an online stage where individuals can collaborate with each other for their own or expert interests. There is a major issue of Spam in Social Media where Spam is unimportant or spontaneous messages sent over the Internet, commonly to countless, for the reasons for spreading malware, promoting, and phishing and so on. In this paper, YouTube Spam issue is examined as should be obvious that there are various spam remarks on YouTube which do not have any importance to a specific post or video. To examine a gigantic measure of the dataset, a computerized apparatus is required which is administered by Machine Learning where Machine learning is a kind of Artificial Intelligence (AI) that enables programming applications to wind up more exact in anticipating results without being unequivocally customized. The fundamental start of machine learning is to manufacture calculations that can get input information and utilize measurable investigation to anticipate a yield an incentive inside an adequate range. Machine learning calculations are regularly ordered as being supervised or unsupervised. The dataset is examined based on classification technique of supervised learning. In this interest, different popular existing algorithms are compared and another algorithm is produced utilizing ensembling approach. The best precision accomplished is 76.3%.

Index Terms- Social Media, YouTube, Artificial Intelligence, Machine Learning, Ensembling Approach

I. INTRODUCTION

People require some dialect to converse with other individuals. Furthermore, for that, it is fundamental that same dialect ought to be known by both the gatherings. What's more, in that, it is likewise known how to recognize words in light of tone, setting, and so on. It can be comprehended when somebody is talking snidely, making a joke, sincerely tormented, or is likely a spambot. Similarly, a mechanized device is required which can foresee these highlights in Social Media. For this, machine learning is utilized.

Machine learning is a field of Artificial Intelligence (AI) that enables programming applications to wind up more exact in anticipating results without being explicitly modified. The essential commence of machine learning is to assemble calculations that can get input information and utilize factual investigation to anticipate a yield an incentive inside a satisfactory range. Machine learning calculations are frequently arranged as being supervised or unsupervised.

Web-based social networking is a progression of sites and applications intended to enable individuals to share content rapidly, productively and progressively. Different well-known person to person communication destinations are Facebook, Twitter, YouTube, LinkedIn, Instagram and so on.

The sheer volume of web-based social networking movement requires mechanized devices to manage the handling exercises. It is inconceivable, even with a committed online networking group, to monitor all channels and brand noticed. Rather, web scratching instruments accumulate every one of the posts that might be related with the brand, place them in an information lake from which they are bolstered into the calculations that cut up them into important pieces. A case can be viewed as distinguishing which remarks to consider as spam in YouTube.

II. LITERATURE SURVEY

a. The Impact of Social Media on the Academic Development of School Students

Today, it is urgent to decide the effect of web-based social networking on the scholarly execution of understudies. The discoveries show that there is no connection between online networking and scholarly execution; this is plainly anticipated in their general review normal. As saw in the discourse, regardless of whether the understudies spend short of what one hour via web-based networking media or over six hours via

web-based networking media, or even the normal measure of time which is between one to three and three to six hours every day, understudies still offer a similar review extend normally. 61% of the respondents have the most noteworthy review ranges which are 90%-100% and they changed between each of the four time runs via web-based networking media every day. By this, it is protected to infer that there is no negative effect on the utilization of online networking on the scholastic execution of the school understudies.

b. From Social Media to Public Health Surveillance: Word Embedding based Clustering Method for Twitter Classification

Life fulfilment alludes to a to some degree stable psychological appraisal of one's own life. The other part is influenced: the harmony between the nearness of positive and negative feelings in everyday life. While impact has been contemplated utilizing online networking datasets (especially from Twitter), life fulfilment has gotten next to zero consideration. Here, they look at patterns in posts about existence fulfilment from a two-year test of Twitter information. They applied an observation procedure to remove articulations of both fulfilment and disappointment with life. An important outcome is that predictable with their definitions drifts in life fulfilment presents are insusceptible on outside occasions (political, regular and so forth.) not at all like influence patterns announced by past scientists. Looking at clients they discovered contrasts amongst fulfilled and disappointed clients in a few semantic, psychosocial and different highlights. For instance, the last post more tweets communicating outrage, nervousness, wretchedness, trouble and on death. They likewise think about clients who change their status after some time from happy with life to disappointed or the other way around. Imperative is that the psychosocial tweet highlights of clients who change from fulfilled to disappointed are very unique in relation to the individuals who remain fulfilled after some time. Generally speaking, the perceptions they made are steady with instinct and reliable with perceptions in the sociology explore.

c. Life Satisfaction and the Pursuit of Happiness on Twitter

Online networking gives a minimal effort elective hotspot for general wellbeing reconnaissance and wellbeing related order assumes an essential part to recognize helpful data. This paper abridged the current arrangement techniques utilizing online networking in general wellbeing. These techniques depend on pack of-words (BOW) display and experience issues getting a handle on the semantic significance of writings. Dissimilar to these techniques, they exhibit a word installing based grouping strategy. Word inserting is one of the most grounded slants in Natural Language Processing (NLP) as of now. It takes in the ideal vectors from encompassing words and the vectors can speak to the semantic data of words. A tweet can be spoken to as a couple of vectors and isolated into bunches of comparable words. As per closeness measures of the considerable number of groups, the tweet would then be able to be named related or irrelevant to a subject (e.g., flu). Our reproductions demonstrate a decent execution and the best precision accomplished was 87.1%. In addition, the proposed strategy is unsupervised. It doesn't expect work to name preparing information and can be promptly stretched out to other characterization issues or other diseases. In this pursuit, they outline the current order approaches in general wellbeing. Once the fleeting illness related tweets are gathered through the proposed arrangement strategy, they can utilize remove based exceptions technique for distinguishing episodes.

III. DATASET DESCRIPTION

Dataset is gathered from UCI Repository which has additionally gathered the corpus from YouTube Data API v3. It is an open arrangement of remarks gathered for spam examine. It has five datasets made out of 1956 genuine messages removed from five recordings that were among the 10 most saw on the gathering time frame.

Data Set Characteristics:	Text	Number of Instances:	1956	Area:	Computer
Attribute Characteristics:	N/A	Number of Attributes:	5	Date Donated	2017-03-26
Associated Tasks:	Classification	Missing Values?	N/A	Number of Web Hits:	35057

Figure 1. Dataset Information

The table underneath records the YouTube video ID, the datasets, the quantity of tests in each class and the aggregate number of tests per dataset.

Table 1. Dataset Description

Dataset	YouTube ID	Spam	Ham	Total
LMFAO	KQ6zr6kCPj8	236	202	438
Shakira	pRpeEdMmmQ0	174	196	370
KatyPerry	CevxZvSJLk8	175	175	350
Psy	9bZkp7q19f0	175	175	350
Eminem	uelHwf8o7_U	245	203	448

The gathering is made out of one CSV document for every dataset, where each line has the accompanying attributes:

AUTHOR , COMMENT_ID, CONTENT , DATE, CLASS where COMMENT_ID tells that remark is done from which ID, AUTHOR tells the name of the individual, DATE tells on which date remark is posted, CONTENT tells what was the substance of the remark and CLASS credit is utilized to group information whether it is spam or not, utilizing Classification algorithm. '0' implies that information isn't spammed and is significant to the specific video and '1' means that information is spammed or isn't identified with the video.

IV. DATA FLOW

At first, data is collected from UCI Repository and then data cleansing is performed to remove redundancy and missing values in dataset. In feature selection, minimum set of data is selected that gives the best performance. Then data is divided for training and testing purpose. Best and general rule to divide data is 70-30 but can also be divided to 60-40, 50-50 and many more combinations. Training data is used to perform various computations and testing data is used to compare with the existing ones. Cross validation is used to check the robustness of the model. For this, k-fold validation is used and the value of k used is 10. And at the end, result is analyzed.

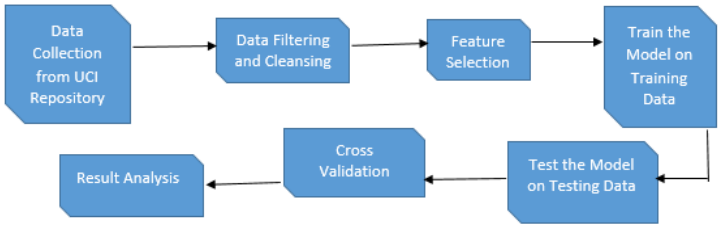

Figure 2: Data Flow

V. METHODOLOGY

a. Algorithms

5.1.1 AdaBoost Classification Trees Algorithm: AdaBoost, shortly known for Adaptive Boosting, is a machine learning meta-calculation. It can be utilized as a part of conjunction with numerous different sorts of learning calculations to enhance execution. AdaBoost is touchy to uproarious information and anomalies. In a few issues, it can be less helpless to the over fitting issue than other learning calculations. The individual students can be powerless, however as long as the execution of every one is marginally superior to anything arbitrary speculating, the last model can be demonstrated to merge to a solid student.

5.1.2 Boosted Classification Trees: Boosting is a machine learning outfit meta-calculation for fundamentally lessening inclination, and furthermore fluctuation in administered learning, and a group of machine learning calculations that change over frail students to solid ones. A feeble student is characterized to be a classifier that is just somewhat associated with the genuine arrangement (it can mark illustrations superior to anything arbitrary speculating). Conversely, a solid student is a classifier that is discretionarily very much connected with the genuine arrangement.

5.1.3 Least Square Support Vector Machine with Radial Basis Function Kernel: Slightest squares bolster vector machines (LS-SVM) are minimum squares renditions of help vector machines (SVM), which are an arrangement of related regulated learning strategies that break down information and perceive examples, and which are utilized for characterization and relapse investigation. In this variant, one finds the arrangement by understanding an arrangement of direct conditions rather than a curved quadratic

programming (QP) issue for established SVMs. LS-SVMs are a class of bit based learning strategies.

b. R Tool

R is a dialect and condition for measurable figuring and illustrations. It is a GNU venture which is like the S dialect and condition which was produced at Bell Laboratories (some time ago AT&T, now Lucent Technologies) by John Chambers and partners. R can be considered as an alternate execution of S. There are some imperative contrasts, yet much code composed for S runs unaltered under R.

c. Prediction of Classification Algorithm and its Parameters

Five algorithms are used to predict the result namely, AdaBoost Classification Trees, Boosted Classification Trees, Least Square Support Vector Machine with Radial Basis Function Kernel, Conditional Inference Tree and Multivariate Adaptive Regression Spline. The computations performed on these five algorithms are shown in the table below:

Table 2: Algorithm Classification

Algorithm Name	Libraries	Package	Precision	Recall	F1 Square	Accuracy
AdaBoost Classification Trees	fastAdaBoost	caret	0.684	0.715	0.699	78.93
Boosted Classification Trees	ada, plyr	caret	0.657	0.751	0.701	85.78
Least Square Support Vector Machine with Radial Basis Function Kernel	kernlab	Caret	0.672	0.728	0.699	79.20
Conditional Inference Tree	Party	Caret	0.719	0.668	0.692	69.23
Multivariate Adaptive Regression Spline	Earth	caret	0.625	0.788	0.697	70.03

In AdaBoost Classification Trees, the accuracy achieved is 78.93%. In Boosted Classification Trees, the accuracy achieved is 85.78%. In Least Square Support Vector Machine with Radial Basis Function Kernel, the accuracy achieved is 79.20%. In Conditional Inference Tree, the accuracy achieved is 69.23. In Multivariate Adaptive Regression Spline, the accuracy achieved is 70.03.

The following parameters are shown with graph as:

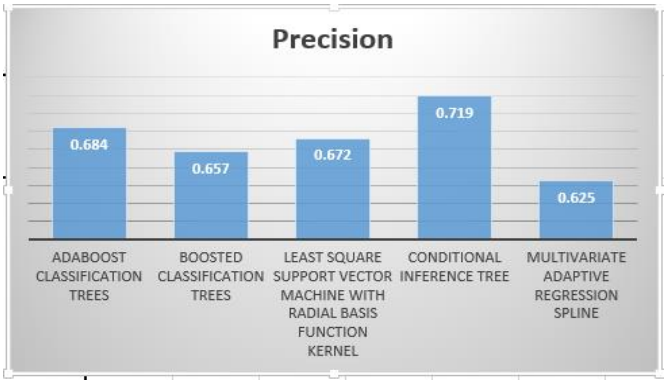

Figure 3: Precision

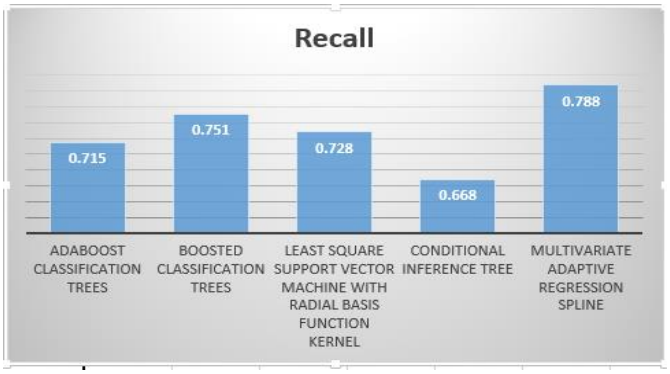

Figure 4: Recall

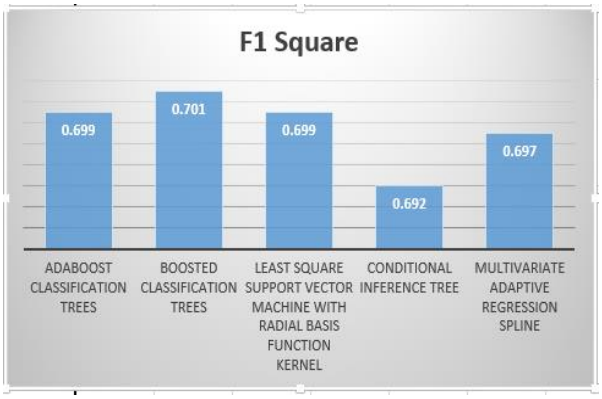

Figure 5: F1 Square

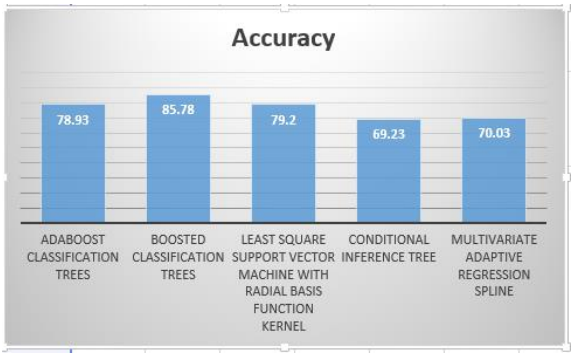

Figure 6: Accuracy

From these, best three algorithms are chosen and shown in the table below:

Table 3: Best Three Algorithms on the Basis of Accuracy

Algorithm Name	Accuracy
AdaBoost Classification Trees	78.93
Boosted Classification Trees	85.78
Least Square Support Vector Machine with Radial Basis Function Kernel	79.20

These best 3 algorithms are shown graphically as:

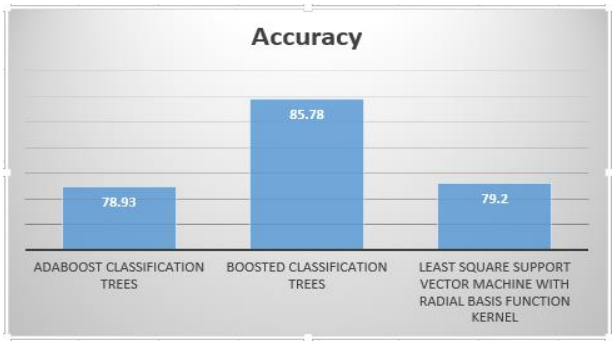

Figure 7: Best 3 Algorithms on the Basis of Accuracy

VI. RESULT

a. Ensembling Approach 10-fold Method

i. Ensembled Algorithm:

1. Choose best three existing model on the basis of highest accuracy.
2. Now, combine the prediction of best model using caretEnsemble package.
3. Run the ensemble model.
4. Apply 10-fold cross validation method.
5. Result
6. Exit

Using ensembling approach, a new algorithm is generated and is executed 10 times. Its precision, recall, F1 Square and accuracy is note down below:

Table 4: 10-fold Method

Runs	Precision	Recall	F1 Square	Accuracy
1	0.668	0.735	0.700	84.9
2	0.733	0.648	0.688	87.9
3	0.625	0.788	0.697	83.6
4	0.684	0.715	0.699	85.2
5	0.657	0.751	0.701	84.6
6	0.647	0.763	0.700	84.3
7	0.679	0.720	0.699	85.1
8	0.719	0.668	0.692	85.8
9	0.672	0.728	0.699	85.0
10	0.636	0.772	0.697	83.9

These parameters can be shown graphically as:

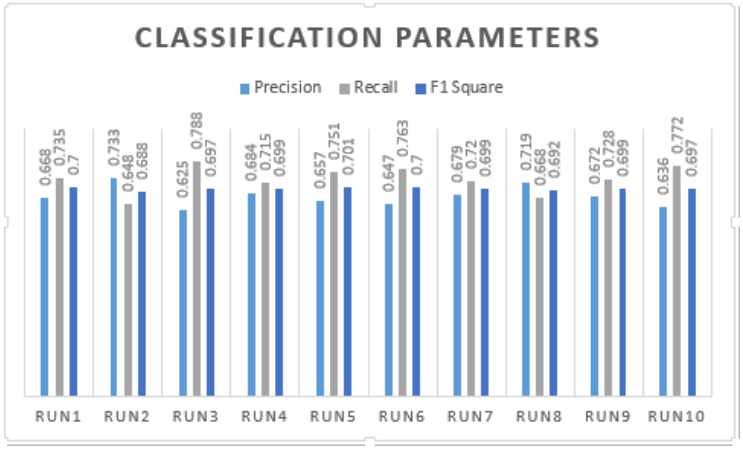

Figure 8: Classification Parameters

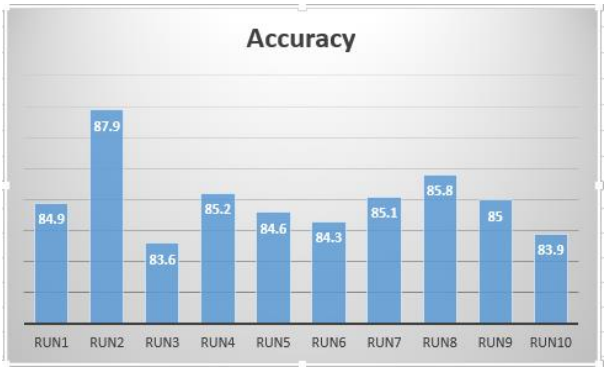

Figure 9: Best Accuracy Achieved

After applying ensemble approach, the algorithm generated by this research will always give accurate result for YouTube Spam Collection dataset. 10-fold cross validation method used here to give highest accuracy i.e. 87.9%.

VII. CONCLUSION

In this paper, YouTube Spam issue is examined as should be obvious that there are various spam remarks on YouTube which do not have any importance to a specific post or video. To examine a gigantic measure of the dataset, this research predict the accuracy of various existing algorithms such as in AdaBoost Classification Trees, the accuracy achieved is 78.93%, in Boosted Classification Trees, the accuracy achieved is 85.78%, in Least Square Support Vector Machine with Radial Basis Function Kernel, the accuracy achieved is 79.20%, in Conditional Inference Tree, the accuracy achieved is 69.23 and in Multivariate Adaptive Regression Spline, the accuracy achieved is 70.03. After this, an ensembling approach is applied to generate a new algorithm and from that a higher accuracy i.e. 87.9% is achieved which is more than the accuracy of existing algorithms.

ACKNOWLWDGEMENT

With immense please I, Ms. Shikha Kaushik presenting "Predicting Human Well-Being using Social Media via Machine Learning" Seminar presentation as part of the curriculum of Bachelor Degree. I wish to thank all the people who gave me unending support.

I express my profound thanks to seminar coordinator Ms. Shalini Puri.

I am thankful to my seminar guide Ms.Prachi Sharma for her kind support and providing me expertise of the domain to develop the research paper.

REFERENCES

[1] Yang C, Srinivasan P. Life satisfaction and the pursuit of happiness on Twitter. PloS one. 2016 Mar 16;11(3):e0150881.

[2] Dai X, Bikdash M, Meyer B. From social media to public health surveillance: Word embedding based clustering method for twitter classification. InSoutheastCon, 2017 2017 Mar 30 (pp. 1-7). IEEE.

[3] El-Badawy TA, Hashem Y. The impact of social media on the academic development of school students. International Journal of Business Administration. 2014 Dec 16;6(1):46.

[4] Manago AM, Vaughn L. Social media, friendship, and happiness in the millennial generation. InFriendship and Happiness 2015 (pp. 187-206). Springer, Dordrecht.

[5] Lee K, Caverlee J, Webb S. Uncovering social spammers: social honeypots+ machine learning. InProceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval 2010 Jul 19 (pp. 435-442). ACM.

[6] https://www.slideshare.net/pacoid/data-workflows-for-machine-learning

[7] https://www.thebalance.com/what-is-social-media-2890301

[8] https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection

[9] https://en.wikipedia.org/wiki/AdaBoost

[10] https://en.wikipedia.org/wiki/Least_squares_support_vector_machine

[11] https://www.r-project.org/about.html