AN INNOVATIVE METHOD TO ENHANCE THE ACCURACY IN THE CLASSIFICATION OF SPAM DETECTION FOR YOUTUBE COMMENTS WITH USING LOGISTIC REGRESSION OVER K–NEAREST NEIGHBOR

S. Venkata Manoj Kumar Reddy[1], K. Malathi [2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: The aim of this research article is to detect the spam comments on YouTube videos with an enhanced accuracy rate by using Innovative Logistic Regression (LR) in comparison with K-Nearest Neighbor (KNN) Classifier.

Materials & Methods: The data set in this paper utilizes UCI machine learning repositories. The sample size of predicting the spam comments on YouTube videos with enhanced accuracy rate was sample 80 (Group 1=40 and Group 2 =40) and calculation is performed utilizing G-power 0.8 with alpha and beta qualities are 0.05, 0.2 with a confidence interval at 95%. Predicting the spam comments on YouTube videos with enhanced accuracy rate is performed by Innovative Logistic Regression (LR) whereas a number of samples (N=10) and K-Nearest Neighbor (KNN) were the number of samples (N=10).

Results: The Innovative Logistic Regression (LR) classifier has 94.4785 higher accuracy rates when compared to the accuracy rate of K-Nearest Neighbor (KNN) is 88.3435. The study has a significance value of $p<0.05$ i.e. $p=0.0243$.

Conclusion: Innovative Logistic Regression (LR) provides the better outcomes in accuracy rate when compared to K-Nearest Neighbor (KNN) for detecting the spam comments on YouTube videos with enhanced accuracy rate.

Keywords: Youtube Spam Comment Detection, Innovative Logistic Regression Classifier, K-Nearest Neighbor Classifier, Accuracy Rate, Machine Learning.

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105.

[2*]Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105.

1. Introduction

The goal of the research article is to increase the accuracy of predicting spam comments on YouTube videos. YouTube is one of the most famous and well-known social media sites. YouTube is functioning as for the user to upload or share any relevant videos. Any Internet user from all over the world can watch the video online. From the video in YouTube, users not only can share their videos, but also can comment on the videos. Comments that came from the users sometimes not only to praise the good video or criticize videos they dislike but also post an unwanted or unsolicited and unrelated electronic message that is sent in bulk to a group of recipients which is also known as spam(Stukus, Patrick, and Nuss 2019). Spam causes many problems, including wasting the user's time, memory and using up network bandwidths. Organizations and users could face financial loss due to the threat of spam(Stukus, Patrick, and Nuss 2019). Some of the spammers use the comment part on YouTube for advertising issues, while others are responsible for distributing computer viruses and there are some spam messages intended to steal the user's financial identities(Council of Europe 2007). The most concerning threats of spam are when involving malicious spam that will lead to phishing websites once the users click the link(Aiyar and Shetty 2018) and the distribution of malware. To overcome this problem and to protect the online social networks from spammers and promoters, we proposed a new approach to classify the users as legitimate, spammers, or promoters by using an Innovative logistic regression algorithm. The statistical analysis of results indicates that the proposed logistic regression method shows good accuracy results.

Recently, many techniques have been proposed for automatic spam comment detection that can be categorized into machine learning (ML) and deep learning (DL) techniques based on the feature selection and learning mechanism. IEEE Explore published 117 research papers, and Google Scholar found 158 articles. Lee et. al propose a multi-level hierarchical system to identify offensive videos. They use multimedia content (frame, color, images), contextual metadata, hash signature (encrypted video header) as discriminatory features (Lee, Shim, and Kim 2009). proposed a framework to detect Web spamming which uses social network mettrics. A framework to detect spamming in tagging systems, which is a type of attack that aims at raising the visibility of specific objects, was proposed in (Harth and Koch 2012). Chowdhury Rashid et al. (Rahman et al. 2020) generated a lift chart by using three different data mining models. This lift chart finds out the lift score when compared to a random guess. The predicted probability for the three different data mining models i.e. Naïve Bayes, Decision Tree, and Clustering is calculated. Tulio C. Alberto et al.(Alberto, Lochter, and Almeida 2015) used different classification algorithms i.e. Naïve Bayes, Decision tree, SVM, Random forest, and Innovative logistic regression on five different datasets. They have achieved a confidence level of almost 99 % on all these classifiers. Saumya Goyal et al. (Goyal, Chauhan, and Parveen 2016) spam message detection on real twitter social media dataset is applied using KNN and decision tree. A study conducted by Alberto et al., (Alberto, Lochter, and Almeida 2015) were using six (6) classifiers to find which classifier gives better performance in detecting YouTube spam comments. Classifier users consist of K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), and Innovative Logistic Regression (LR). The most cited article was (Alberto, Lochter, and Almeida 2015; Othman and Din 2019) in IEEE Explore with 19 citations and 785 full text views.

Our team has extensive knowledge and research experience that has translated into high quality publications (K. Mohan et al. 2022; Vivek et al. 2022; Sathish et al. 2022; Kotteeswaran et al. 2022; Yaashikaa, Keerthana Devi, and Senthil Kumar 2022; Yaashikaa, Senthil Kumar, and Karishma 2022; Saravanan et al. 2022; Jayabal et al. 2022; Krishnan et al. 2022; Jayakodi et al. 2022; H. Mohan et al. 2022). The main drawback with this existing method is in large datasets, this can lead to the model being overly matched in the training set, which means exceeding the accuracy of the predictions in the training set and the model may not be able to predict accurate results in the test set. This paper proposed a Innovative logistic regression (LR) classifier to detect the spam comments on YouTube videos and compares the results with the K-Nearest Neighbor (KNN) classifier on the basis of accuracy, precision, and recall values. Innovative Logistic regression is easier to train and implement as compared to the KNN method. The aim of this paper is to evaluate the accuracy of the LR and K-NN classifier before and after optimizing the important parameters using the Python software tool. The comparison of the two different models LR and K-NN are to be tested. The performance analysis of the proposed spam comment detection method gives better results than the existing K-NN method.

2. Materials and Methods

This work was carried out in the Digital Image Processing Laboratory, Department of Computer Science and Engineering, Saveetha School of Engineering. In this paper, The YouTube Spam Collection Data Set Collect from UCI machine learning repositories(Lichman and Others 2013). The dataset contained ten selected videos and were downloaded from YouTube through API. It is composed of 1,900 real and non-encoded messages that were labeled as legitimate (ham) or spam. Each sample represents a text comment posted in the comments section of each selected video. Group 1 was a K-Nearest Neighbor (KNN) algorithm and Group 2 was an innovative Logistic regression (LR) model. Python 3.9.2 and the NLP library were used for all implementations. The calculation is performed utilizing G-power 0.85 with alpha and beta qualities 0.05, 0.1 with a confidence interval at 95%.

K-Nearest Neighbor

K-NN is a supervised learning method. Data is appearing in a vector space in the K-NN algorithm. K–NN emphasizes k most similar training data points to a testing data point. After determining the K-Nearest Neighbors, the algorithm will combine the neighbors' to decide the label of the testing data point. The basics of the format is as follows: the quality of the space, and the majority of the samples in the k-Nearest neighbors of a sample to be classified to belong to one category, this is the definition of a sample to be classified is that the selection is also a part of the category of. The main ideas of the K-NN classification, the algorithm is as follows: given an unknown sample, the distance between the training sample and the unknown sample is calculated, and then, as the k-nearest samples are selected from a set of training samples, the shortest distance between the training sample and the sample. The classification of unknown samples is carried out according to the category of k-nearest samples. In the algorithm, it is assumed that all of the samples correspond to points in the n dimensional space Rn. In an n-dimensional space, the set of feature vectors for the sample of x looks something like this:

$$<a1(x),a2(x)\ldots an(x)> \quad (1)$$

The distance between the two samples x_ and x_ is indicated as (,) *i j d x x* , Expressed by Euclidean distance is as follows

In accordance with the following formula: (2), the k-nearest neighbor of the ranked sample is to be selected. The most important points that play an important role in the implementation of the K-NN algorithmic model are as follows:

(1) The training of sampling;
(2) The Mathematical model for the calculation of the distance contract.
(3) The selection of the K value.
(4) The basis for the classification;

This is key in the above-mentioned points will have a significant impact on the classification accuracy of K-NN algorithm. The classification process consists of the symptoms that you do not have an impact on the classification of, or to have a low impact on the rating, reducing the accuracy of the classification.

The sample group 1 is the K-Nearest Neighbor (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. The steps involved in the implementation of the KNN algorithm are described as follows.

Logistic Regression

Innovative Logistic regression used as a statistical model for finding the probability of a certain class in terms of discrete or categorical results such as yes or no, 1 or 0. It can be binomial, ordinal, or multidimensional. This is used for measuring the relationship between categorical dependent variables and one or more independent variables by estimating probability using a logistic function. Also very easy to implement, interpret, and very efficient to train. Innovative Logistic regression is used as a classifier as a novel prediction that predicts the result of a reliant variable through a bunch of indicators. It is appropriate where the subordinate variable is all out and dichotomous and autonomous factors are all out, constant or blended. The reliant variable in calculated relapse takes the worth of 1 (one) with likelihood of achievement of an occasion, or the worth of 0 (zero) with the likelihood of failure of an occasion.

Least substantial example size needed for LR Model:

Number of cases N= (10 k (predictor Variable)) / (Cases(p)) (2)

Where,
N = Minimum example size needed for model
k = Number of free/Predictor variable
p = the littlest of the extents of negative or positive cases.

From the aftereffects of Innovative Logistic Regression demonstrating, the co-productive are tried for importance utilizing a few tests like Wald Chi Square Test, Likelihood-Ratio Test, and Deviance test to approve the model. A Wald Chi-square test is utilized to test whether at least two factors are free or homogeneous. The chi-square test for autonomy looks at whether knowing the worth of one variable assists with assessing the worth of another variable. The probability proportion test utilizes the proportion of the expanded worth of the probability work for the full model over the boosted worth of the probability work for the basic model. It is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of Innovative logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent predictor or explanatory variables. The sample preparation group 2 is the Innovative Logistic regression (LR) algorithm, which is one of the machine learning algorithms used for solving classification problems. It is used to estimate probability whether an instance belongs to a class or not. The experimental results show that the proposed LR method has achieved better accuracy results.

Statistical Analysis

The output is obtained by using Python software. To train these datasets, required a monitor with resolution of 1024×768 pixels (7th gen, i5, 4 8GB RAM, 500 GB HDD), and Python software with required library functions and tool functions. For statistical implementation, the software tool used here is IBM SPSS V26.0(Hilbe 2004). The independent sample t test was performed to find the mean, standard deviation and the standard error mean statistical significance between the groups, and then comparison of the two groups with the SPSS software will give the accurate values for the two different s which will be utilized with the graph to calculate the significant value with maximum accuracy value (94.47), mean value (94%) and standard deviation value (0.43733). Dependent variables are accuracy and independent variables are KNN and Logistic Regression (LR) methods.

3. Results

Figure.3. shows the simple bar graph for K-Nearest Neighbor (KNN) Classifier accuracy rate is compared with Innovative Logistic Regression (LR) Classifier. The Innovative Logistic Regression (LR) Classifier is higher in terms of accuracy rate 94.4785 when compared with K-Nearest Neighbor (KNN) Classifier 88.3435. Variable results with its standard deviation ranging from 80 lower to 90 higher K-Nearest Neighbor (KNN) Classifier where Logistic Regression (LR) Classifier standard deviation ranging from 90 lower to 100 higher. There is a significant difference between the K-Nearest Neighbor (KNN) Classifier and the Innovative Logistic Regression (LR) Classifier (p<0.05 Independent sample test). X-axis: Logistic Regression (LR) Classifier accuracy rate vs K-Nearest Neighbor (KNN) Classifier Y-axis: Mean of accuracy rate, for identification of keywords ± 1 SD with 95 % CI. Table.1 shows the Evaluation Metrics of Comparison of K-Nearest Neighbor (KNN) and Logistic Regression (LR) Classifier. The accuracy rate of K-Nearest Neighbor (KNN) is 88.3435 and Logistic Regression (LR) is 94.4785. In all aspects of parameters Logistic Regression (LR) provides better performance compared with the K-Nearest Neighbor (KNN) of predicting Spam comments on YouTube videos with improved accuracy rate. Table.2 shows the statistical calculation such as Mean, standard deviation and standard error Mean for K-Nearest Neighbor (KNN) and Logistic Regression (LR). The accuracy rate parameter used in the t-test. The mean accuracy rate of K-Nearest Neighbor (KNN) is 88.3435 and Logistic Regression (LR) is 94.4785. The Standard Deviation of K-Nearest Neighbor (KNN) is 1.01939 and Logistic Regression (LR) is 0.43733. The Standard Error Mean of K-Nearest Neighbor (KNN) is 0.67382 and Innovative Logistic Regression (LR) is 0.12234. Table.3 displays the statistical calculations for independent samples tested between K-Nearest Neighbor (KNN) and Innovative Logistic Regression (LR). The significance for signal to noise ratio is 0.0243. Independent samples T-test is applied for comparison of K-Nearest Neighbor (KNN) and Innovative Logistic Regression (LR) with the confidence interval as 95% and level of significance as 0.33232. This independent sample test consists of significance as 0.001, significance (2-tailed), Mean difference, standard error difference, and lower and upper interval difference.

4. Discussion

In this section, we evaluate our results and also define the evaluation criteria to calculate the performances of our classification models. During the training process, the confusion matrix was used to evaluate the classification models. The confusion matrix is a matrix that maps the predicted outputs across actual outputs. It is often used to describe the performance of a classification model on a set of test data. Important metrics were computed from the confusion matrix in order to evaluate the classification models. In addition to correct classification rate or accuracy other metrics that were computed for evaluation were True Positive Rate (TPR), False Positive Rate (FPR), Precision, Accuracy, F1 score, and Misclassification rate. AC is the proportion of the total number of predictions that were correct. AC is calculated as the number of all correct predictions divided by the total number of all predictions. The true positive rate (TPR) also known as sensitivity or recall is the proportion of positive cases that were correctly identified. TPR or Recall is calculated as the number of correct positive predictions divided by the total number of true positives and false negatives. PR is the proportion of the predicted positive cases that were correct. PR is calculated as the number of correct positive predictions divided by the total number of positive predictions. Also known as the ‘Misclassification rate’ is calculated as the number of all false predictions divided by the total number of all predictions. The concept of anomaly detection wherein the divergence from authentic emails was used as a metric to classify emails as spam or ham. Better accuracy was achieved owing to the limited training sets as seen in labeling based systems(Santos et al. 2011). Another common alternative is automatic blocking spammers users that disseminate spam(Benevenuto et al. 2009; Campanha, Lochter, and Almeida 2014). However, unlike spam disseminated in other social networks and email(Almeida and Yamakami 2012; Wang, Irani, and Pu 2011), the spam posted on YouTube is not usually created by bots, but posted by real users aiming for self-promotion on popular videos. Therefore, such messages are more difficult to identify due to its similarity to legitimate messages. In this paper, the author presents an online spam filtering system that can be used in real-time to inspect messages generated by users. The author applies this technique to support the online spam detection problem with sufficiently low expenses. It drops messages classified as “spam” before they reach the recipients, thus protecting them from various kinds of fraud. In this research, the development of a spam comment detection framework by using machine learning techniques has been done. It is important to improve security since the Internet nowadays indicates the security issues(Salleh et al. 2018). There are many studies aimed to reduce attack and to protect user privacy but yet lacking in applying the techniques for social media(Umapathy and Khare 2018). This paper also wants to contribute by examining the suitable features based on the real comment from social media sites for developing spam comment detection framework. There are several phases involved in the development of this framework such as Data Collection, Pre-processing, Features Selection and Extraction, Classification and Detection. Each of these phases has been validated through experiments by using machine learning techniques. The Data Collection is downloaded from UCI Machine Learning and the Pre-processing will clean the dataset before the experiments are performed. However, the proposed method was found inadequate in analyzing the feelings and opinions expressed in platforms such as YouTube. For future research more data would be collected as well as more classification methods could be used in order to get accurate results.

5. Conclusion

In this research, the development of a Youtube spam comment detection framework by using machine learning techniques has been done. It is important to improve security since the Internet nowadays indicates the security issues. The proposed model exhibits the K-Nearest Neighbor (KNN) and Innovative Logistic Regression (LR), in which the Innovative Logistic Regression (LR) has the highest values. The accuracy Rate of Innovative Logistic Regression (LR) is 94.4785 is higher compared with K-Nearest Neighbor (KNN) that has an accuracy rate of 88.3435 in analysis of detecting Spam comments on YouTube videos with improved accuracy rate.

Declarations

Conflicts of Interest

No conflict of interest in this manuscript

Authors Contributions

Author KR was involved in data collection, data analysis & manuscript writing. Author MVP was involved in conceptualization, data validation, and critical review of manuscripts.

Acknowledgment

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences (Formerly known as Saveetha University) for successfully carrying out this work.

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. Chipontime Technologies Pvt. Ltd. Bangalore.
2. Saveetha University
3. Saveetha Institute of Medical And Technical Sciences
4. Saveetha School of Engineering

6. References

Aiyar, Shreyas, and Nisha P. Shetty. 2018. “N-Gram Assisted Youtube Spam Comment Detection.” Procedia Computer Science. https://doi.org/10.1016/j.procs.2018.05.181.

Alberto, Tulio C., Johannes V. Lochter, and Tiago A. Almeida. 2015. “TubeSpam: Comment Spam Filtering on YouTube.” 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). https://doi.org/10.1109/icmla.2015.37.

Almeida, Tiago A., and Akebo Yamakami. 2012. “Occam’s Razor-Based Spam Filter.” Journal of Internet Services and Applications 3 (3): 245–53.

Benevenuto, Fabrício, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, and Marcos Gonçalves. 2009. “Detecting Spammers and Content Promoters in Online Video Social Networks.” In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 620–27. SIGIR ’09. New York, NY, USA: Association for Computing Machinery.

Campanha, Jhony M., Johannes V. Lochter, and Tiago A. Almeida. 2014. “Detecc Ao Automatica de Spammers Em Redes Sociais.” Anais Do XI Encontro Nacional de Intelig ˆencia Artificial E Computacional (ENIAC’14), Sao Carlos, Brazil. http://www.dt.fee.unicamp.br/~tiago/papers/JMC_ENIAC14.pdf.

Council of Europe. 2007. Cyberterrorism: The Use of the Internet for Terrorist Purposes. Council of Europe.

Harth, Andreas, and Nora Koch. 2012. Current Trends in Web Engineering: Workshops, Doctoral Symposium, and Tutorials, Held at ICWE 2011, Paphos, Cyprus, June 20-21, 2011. Revised Selected Papers. Springer Science & Business Media.

Hilbe, Joseph M. 2004. “A Review of SPSS 12.01, Part 2.” The American Statistician 58 (2): 168–71.

Jayabal, Ravikumar, Sekar Subramani, Damodharan Dillikannan, Yuvarajan Devarajan, Lakshmanan Thangavelu, Mukilarasan Nedunchezhiyan, Gopal Kaliyaperumal, and Melvin Victor De Poures. 2022. “Multi-Objective Optimization of Performance and Emission Characteristics of a CRDI Diesel Engine Fueled with Sapota Methyl Ester/diesel Blends.” Energy. https://doi.org/10.1016/j.energy.2022.123709.

Jayakodi, Santhoshkumar, Rajeshkumar Shanmugam, Bader O. Almutairi, Mikhlid H. Almutairi, Shahid Mahboob, M. R. Kavipriya, Ramesh Gandusekar, Marcello Nicoletti, and Marimuthu Govindarajan. 2022. “Azadirachta Indica-Wrapped Copper Oxide Nanoparticles as a Novel Functional Material in Cardiomyocyte Cells: An Ecotoxicity Assessment on the Embryonic Development of Danio Rerio.” Environmental Research 212 (Pt A): 113153.

Kotteeswaran, C., Indrajit Patra, Regonda Nagaraju, D. Sungeetha, Bapayya Naidu Kommula, Yousef Methkal Abd Algani, S. Murugavalli, and B. Kiran Bala. 2022. “Autonomous Detection of Malevolent Nodes Using Secure Heterogeneous Cluster Protocol.” Computers and Electrical Engineering. https://doi.org/10.1016/j.compeleceng.2022.107902.

Krishnan, Anbarasu, Duraisami Dhamodharan, Thanigaivel Sundaram, Vickram Sundaram, and Hun-Soo Byun. 2022. “Computational Discovery of Novel Human LMTK3 Inhibitors by High Throughput Virtual Screening Using NCI Database.” Korean Journal of Chemical Engineering. https://doi.org/10.1007/s11814-022-1120-5.

Lichman, Moshe, and Others. 2013. “UCI Machine Learning Repository, 2013.” URL Http://archive. Ics. Uci. Edu/ml 40.

Mohan, Harshavardhan, Sethumathavan Vadivel, Se-Won Lee, Jeong-Muk Lim, Nanh Lovanh, Yool-Jin Park, Taeho Shin, Kamala-Kannan Seralathan, and Byung-Taek Oh. 2022. “Improved Visible-Light-Driven Photocatalytic Removal of Bisphenol A Using V2O5/WO3 Decorated over Zeolite: Degradation Mechanism and Toxicity.” Environmental Research. https://doi.org/10.1016/j.envres.2022.113136.

Mohan, Kannan, Abirami Ramu Ganesan, P. N. Ezhilarasi, Kiran Kumar Kondamareddy, Durairaj Karthick Rajan, Palanivel Sathishkumar, Jayakumar Rajarajeswaran, and Lorenza Conterno. 2022. “Green and Eco-Friendly Approaches for the Extraction of Chitin and Chitosan: A Review.” Carbohydrate Polymers 287 (July): 119349.

Othman, N. F., and W. Din. 2019. “Youtube Spam Detection Framework Using Naive Bayes and Logistic Regression.” Indonesian Journal of Electrical Engineering and. https://core.ac.uk/download/pdf/220098744.pdf.

Rahman, M. O., A. S. Islam, M. S. Choudhury, A. A. Raihan, M. S. Alam, M. Chowdury, and A. Islam. 2020. “A Study of Association between

H. Pylori Genotype and Chronic Gastritis." Mymensingh Medical Journal: MMJ 29 (3): 664–75.

Salleh, Siti Norussaadah Mohd, Roshidi Din, Nur Haryani Zakaria, and Aida Mustapha. 2018. "A Review on Structured Scheme Representation on Data Security Application." Indonesian Journal of Electrical Engineering and Computer Science. https://doi.org/10.11591/ijeecs.v11.i2.pp733-739.

Santos, Igor, Carlos Laorden, Xabier Ugarte-Pedrero, Borja Sanz, and Pablo G. Bringas. 2011. "Spam Filtering through Anomaly Detection." In E-Business and Telecommunications, 203–16. Communications in Computer and Information Science. Springer, Berlin, Heidelberg.

Saravanan, A., P. Senthil Kumar, B. Ramesh, and S. Srinivasan. 2022. "Removal of Toxic Heavy Metals Using Genetically Engineered Microbes: Molecular Tools, Risk Assessment and Management Strategies." Chemosphere 298 (July): 134341.

Sathish, T., R. Saravanan, V. Vijayan, and S. Dinesh Kumar. 2022. "Investigations on Influences of MWCNT Composite Membranes in Oil Refineries Waste Water Treatment with Taguchi Route." Chemosphere 298 (July): 134265.

Stukus, David R., Michael D. Patrick, and Kathryn E. Nuss. 2019. Social Media for Medical Professionals: Strategies for Successfully Engaging in an Online World. Springer.

Umapathy, Kumaran, and Neelu Khare. 2018. "An Efficient & Secure Content Contribution and Retrieval Content in Online Social Networks Using Level-Level Security Optimization & Content Visualization Algorithm." Indonesian Journal of Electrical Engineering and Computer Science 10 (2): 807–16.

Vivek, J., T. Maridurai, K. Anton Savio Lewise, R. Pandiyarajan, and K. Chandrasekaran. 2022. "Recast Layer Thickness and Residual Stress Analysis for EDD AA8011/h-BN/B4C Composites Using Cryogenically Treated SiC and CFRP Powder-Added Kerosene." Arabian Journal for Science and Engineering. https://doi.org/10.1007/s13369-022-06636-5.

Wang, De, Danesh Irani, and Calton Pu. 2011. "A Social-Spam Detection Framework." In Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference, 46–54. CEAS '11. New York, NY, USA: Association for Computing Machinery.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Algal Biofuels: Technological Perspective on Cultivation, Fuel Extraction and Engineering Genetic Pathway for Enhancing Productivity." Fuel. https://doi.org/10.1016/j.fuel.2022.123814.

Yaashikaa, P. R., P. Senthil Kumar, and S. Karishma. 2022. "Review on Biopolymers and Composites – Evolving Material as Adsorbents in Removal of Environmental Pollutants." Environmental Research. https://doi.org/10.1016/j.envres.2022.113114.

Tables and Figures

Table. 1. Comparison of K-Nearest Neighbor (KNN) and Logistic Regression (LR) Classifier for predicting the spam comments in YouTube videos with improved accuracy rate. The accuracy rate of K-Nearest Neighbor (KNN) is 88.3435 and Logistic Regression (LR) has 94.4785

SI.No.	**Test Size**	**Accuracy Rate**	
		K-Nearest Neighbor	**Logistic Regression**
1	Test 1	84.56	90.67
2	Test2	84.23	90.87
3	Test3	84.32	90.98
4	Test4	85.13	90.45
5	Test5	85.11	91.13

6	Test6	85.55	92.76
7	Test7	86.21	93.14
8	Test8	86.22	93.65
9	Test9	86.34	94.23
10	Test10	87.90	94.24

Table. 2. The statistical calculation such as Mean, standard deviation and standard error Mean for K-Nearest Neighbor (KNN) and Logistic Regression (LR). The accuracy rate parameter used in the t-test. The mean accuracy rate of K-Nearest Neighbor (KNN) is 88.3435 and Logistic Regression (LR) is 94.4785. The Standard Deviation of K-Nearest Neighbor (KNN) is 1.01939 and Logistic Regression (LR) is 0.43733. The Standard Error Mean of K-Nearest Neighbor (KNN) is 0.67382 and Logistic Regression (LR) is 0.12234.

Group		**N**	**Mean**	**Standard Deviation**	**Standard Error Median**
Accuracy	Logistic Regression	10	94.4785	0.43733	0.12234
	K-Nearest Neighbor (Knn)	10	88.3435	1.01939	0.67382

Table. 3. The statistical calculations for independent samples test between K-Nearest Neighbor (KNN) and Logistic Regression (LR). The sig. for signal to noise ratio is 0.0243. Independent samples T-test is applied for comparison of K-Nearest Neighbor (KNN) and Logistic Regression (LR) with the confidence interval as 95% and level of significance as 0.33232. This independent sample test consists of significance as 0.001, significance (2-tailed), Mean difference, standard error difference, and lower and upper interval difference.

Group	**Levene,s Test for Equality of Variances**		**T-Test for Equality of Medians**						
	F	**sig.**	**t**	**df**	**sig(2-tailed)**	**Mean difference**	**Std Error Difference**	**95% Confidence Interval (Lower)**	**95% Confidence Interva1 (Upper)l**

Accuracy	Equal Variances assumed	10.32	0.0243	16.283	18	.001	13.78737	0.64738	12.18924	15.78392
	Equal Variances assumed			12.726	16.457	.001	12.10123	0.10124	11.21283	14.01829

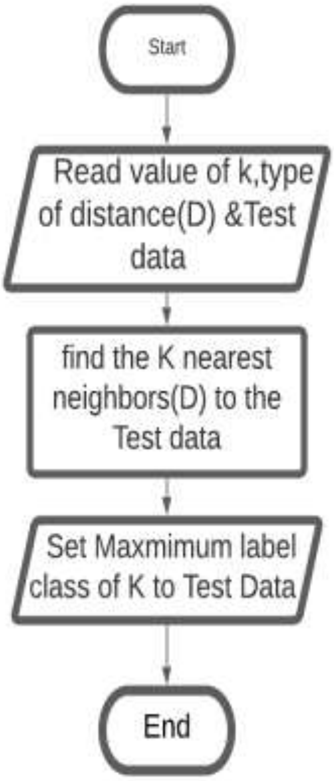
Start
Read value of k,type of distance(D) &Test data
find the K nearest neighbors(D) to the Test data
Set Maxmimum label class of K to Test Data
End

Figure.1: Flow Chart K-Nearest Neighbor (KNN)

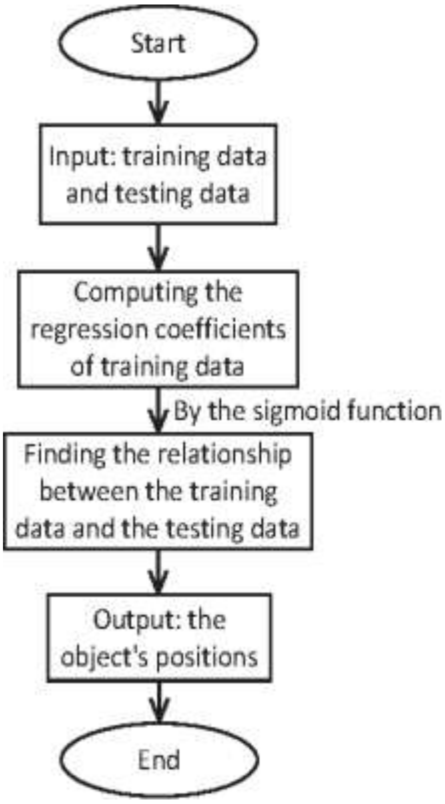
Start
Input: training data and testing data
Computing the regression coefficients of training data
By the sigmoid function
Finding the relationship between the training data and the testing data
Output: the object's positions
End

Figure. 2: Flow chart Logistic regression

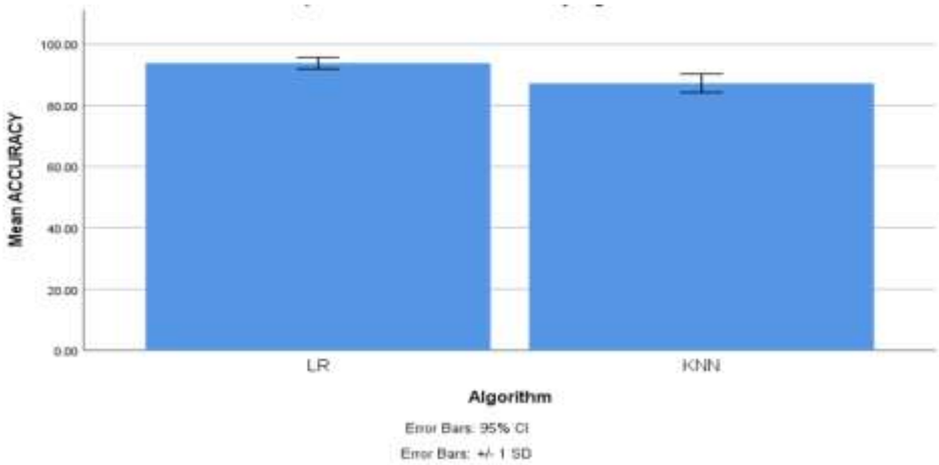

Figure. 3: Bar graph between KNN and Innovative Logistic Regression Classifier. Comparison of KNN algorithm and LR in terms of mean accuracy. The mean accuracy of KNN is better than LR and the standard deviation of KNN is slightly better than LR. X-Axis: KNN vs LR Y-Axis: Mean accuracy of detection ± 1 SD with CI of 95%.