IJMTST is UGC Approved Journal with Journal Id: 43137

International Journal for Modern Trends in Science and Technology, 8(S03):144-151, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST08S0334
Available online at: https://www.ijmtst.com/vol8si03.html

Identification of Malicious Reviewer Groups in Online Product Reviews

Mandapati Rajesh, Dr. RM Pattanayak

Department of Computer Science and Engineering, Godavari Institute of Engineering and Technology (A), JNTUK, Kakinada.

To Cite this Article
Mandapati Rajesh and Dr. RM Pattanayak. Identification of Malicious Reviewer Groups in Online Product Reviews. International Journal for Modern Trends in Science and Technology 2022, 8(S03), pp. 144-151.https://doi.org/10.46501/IJMTST08S0334

Article Info
Received: 26 April 2022; Accepted: 24 May 2022; Published: 30 May 2022.

ABSTRACT

In today's digital world, there is a great deal of opinion spam. It's not unusual for companies to pay people to write favorable or unfavorable reviews of their products. The majority of the time, this is done in a collective context. As a result of previous study, little effort has been made to identify groups that target brands as a whole rather than individual product.Searching an Amazon product review site can identify up to 923 distinct user groups. People who have rated many brands from the same manufacturer are grouped together in "clusters" based on their brand similarity. There are eight characteristics that we believe are unique to each (group or brand) coupling. Use of a supervised classifier to identify radical organizations is described. In order to categories groups based on their members' reviews, a wide range of classifiers are used. The best classifiers are three-layer perceptions. The behavior of such groups can be studied in greater detail in order to better understand brand-level opinion fraud. Consistent ratings, tone of reviews, confirmed purchases, review dates, and helpful votes are all instances of this type of behavior in online reviews. Amazon's current defenses against unauthorized incentives were prompted by a large number of verified reviewers who expressed strong feelings.

Keywords:MRG, Amazon product, radical organizations, Consistent ratings

1.INTRODUCTION

Review sites and online marketplaces now play a major part in purchase decisions. Product with more positive ratings sells better. Purchases increase reviews. Elizabeth Mott [3] Alice [1] believes that increasing sales boosts revenue. In rare circumstances, misleading reviews might sway customers. They can work alone or in groups. However, they can aid other consumers by sharing their own personal experiences. By reviewing something with a sophisticated network, people can have a tremendous impact on how it is viewed (and other approaches outlined in Section VIII). So it works better here. This effect is not limited to biased reviews [2]. 10%-15% of the evaluations are nearly identical to the initial judgment, making a false early review more persuasive than the original judgment itself. Reviews must be aware of and aggressively prevent such conduct. The sales of a product can be influenced and controlled by a large number of people inside an organization's network. For privacy, most ensembles follow a set of rules that are rarely broken. Because they are all owned by the same firm, they have a lot of potential targets. A thorough investigation can help pinpoint these features. New regulation limits the amount of reviews a product can acquire in a single day on Amazon India. Aiming to maintain their efficacy and relevancy, we believe certain organizations target specific brands and examine them rigorously. It can help you gain a competitive advantage

in the highly competitive internet market. Despite many probes, the study into brand-based opinion spamming continues. Brand-related behaviors and consequences must be thoroughly explored to avoid violating review websites' codes of conduct. Why? Unfair market conditions provide certain firms an unfair edge. In this way, non-validated reviewers can access free Supplementary Material Section IV [2 & 3]. This page lists the current radical review organizations. Amazon India uses a second classifier based on reviewer group behavior to detect extreme organizations. Then we can examine the overall trend among these groups and draw conclusions. These are the main conclusions of the studies:

1. In this study, 923 groups of reviewers were manually categorized into "extreme" and "moderate" groups.
2. This was the first time that the problem of recognizing brand-level extremism was described and looked at.
3. The detailed description of extremist review groups.
4. As a way to make our research more likely to be repeatable, we've put the code on GitHub at https://hub.virresh/extremist-reviewers.

The following sections outline the main points of this article. In Section II, we will discuss topics such as review extremism, programmed development based on reviews, and the identification of false reviews, among other things. Section III describes in detail how the data was gathered and how it was annotated, and it concludes with recommendations. Section IV discusses the characteristics and qualities that distinguish a brand [1].

1.1. Research aim and objectives

It is our intention with this programmed to give a new classification method for locating and categorizing online group testimonials of all kinds, which can be used by anyone. This study used a variety of classifiers such as random forest and decision trees, SGD (single-grained decision), logistic regression, naive bayeses, 3 and 4 MLPs, and XG Bost. Based on the characteristics of the reviewer, we are making educated guesses about the outcome [12].

2. RELATED WORK

H. S. Dutta, et al. (2018).Using Hawkes' approach and topic modeling, our team developed Hawkes Eye, a tool for identifying false tweets. Tweets need re-tweets to get traction, which phoney re-tweeters can offer. In order to be considered a "phoney re-tweeter," a person must re-tweet spam, send out a large number of tweets quickly, or use a popular hashtag to promote anything unrelated to the argument. In this paper, we present an updated dataset for detecting false re-tweets. A temporal window is used by HawkesEye rather than an italic>graph-like/italic> connection between tweets or the existence of the entire re-tweeting history of a user to make predictions. S. Dhawan, et al. (2019).As per this study aboutonline fraudster behavior as a whole, it's helpful to examine consumer reviews. Before making a purchase, check out online reviews to get a sense of the product's quality. A sad truth is that spammers use online review boards to fabricate fake customer reviews in an effort to influence sales of specific products. They can be even more damaging than spammers working alone because of the enormous volume of fake reviews they can produce. Spam can be harder to identify at the group level since the concept of a group is nebulous, dynamics within groups change, and data on spam at the group level is scarce. Downloading the Defrauder algorithm we developed to identify groups of online scam reviewers has been completed. Y. Wang, J. Wang, and T. Yao, (2019).Addressed about internet review of it's important to understand what makes for a good internet review so you can write one yourself. A meta analysis of the characteristics of reviews. Through the examination of criteria such as the length, breadth, and timely natures of the reviews, we want to learn more about the elements that influence the value of internet reviews. The researchers conducted a meta-analysis involving 53 empirical studies and 191 effect sizes in order to determine how various features of reviews influenced their findings. When reviews are excessively long or too short in length, they are less useful, however when reviews are well-written, they are more valuable. As an added bonus, we go deep into the website's history and culture in order to find interesting insights about the people who created it. Our findings, which include a number of recommendations, will be of interest to both researchers and online merchants in general. It is critical to provide a quantitative synthesis of this emerging field

of study.Urcuqui, Christian, (2017).Research was done into the use of machine learning to identify malicious websites. We want to make troubleshooting a large Wi-Fi network easier by automating it. We are trying to find out what is causing Wi-Fi networks to decrease in performance by conducting meaningless active scans. Thousands of active scanners are utilized to train a variety of machine learning algorithms. 27 devices from various network setups and providers are used to collect data in a controlled setting. Using unsupervised and supervised machine learning techniques, we found that multilayer perceptions are the most effective model for identifying the causes of active scanning. The in-vivo model validation is done via a Wi-Fi network that is not under our control.A. Almahairi, et al. (2018).As per his analysis of studyIt is beneficial to learn distributed representations from reviews in order to use them for collaborative filtering. As a result of our findings, two more review models have been developed in order to determine how they affect the effectiveness of collaborative filtering. Amazon's expert model data outperforms LDA-based strategies in terms of versatility because of its better adaptability. The recurrent neural network's modeling abilities appear to be interfering with the model's capacity to regularize the representation of products, which is problematic. This is a completely unusual situation.

3. RESEARCH METHODOLOGY

Machine learning (ML) algorithms are a class of automated algorithms that gain knowledge through exposure to real-world data and experience. As far as artificial intelligence is concerned, it's considered part of the package. Machine learning algorithms build a model from training data in order to make predictions or decisions without being explicitly programmed [23]. Traditional algorithms in a variety of fields, such as medicine, email filtering, speech recognition, and computer vision cannot be developed without incorporating principles from AI.

Not all of machine learning is statistical learning, and computational statistics is only a subset of machine-learning. Mathematical optimization provides methods, theories, and application domains in the field of machine learning. An unsupervised learning algorithm is a type of exploratory data analysis algorithm. Some machine learning implementations use data and neural networks in a way that resembles the functioning of a biological brain. In the context of business, predictive analytics is a term for machine learning that is used to predict outcomes [18].The only way to identify them would be to go through the present system and manually identify each and every one of the bogus reviews. The use of machine learning techniques to identify individual users is not an option when trying to identify a huge number of false reviews [13 & 17]. Even when using machine learning algorithms, we are unable to identify each individual user.

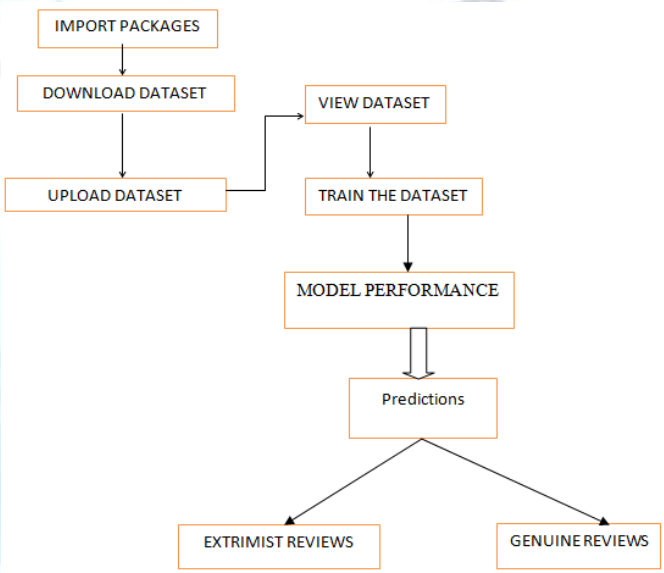


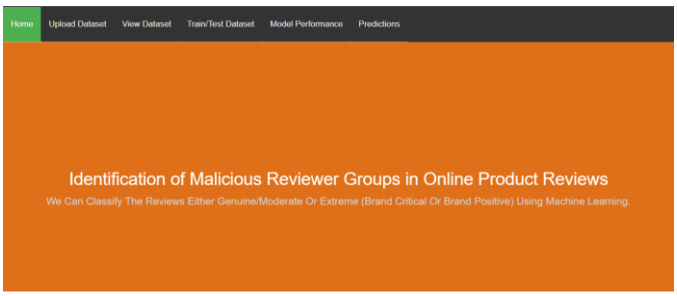
Figure.3.1. Architecture model

3.1. PROPOSED METHOD

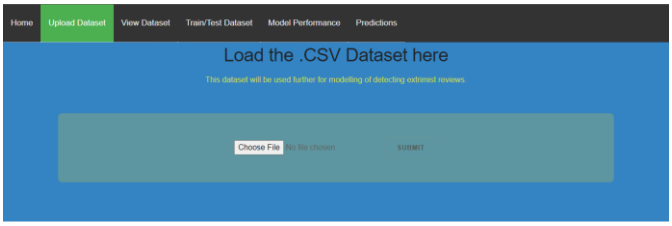
Our proposed strategy, rather than concentrating just on detecting fraudulent reviews, is more concerned with recognizing "radical reviewers," who could or might not be genuine. Furthermore, rather of identifying "individual users," we employ machine learning approaches to identify "groups."

4. RESULT IMPLEMENTATION

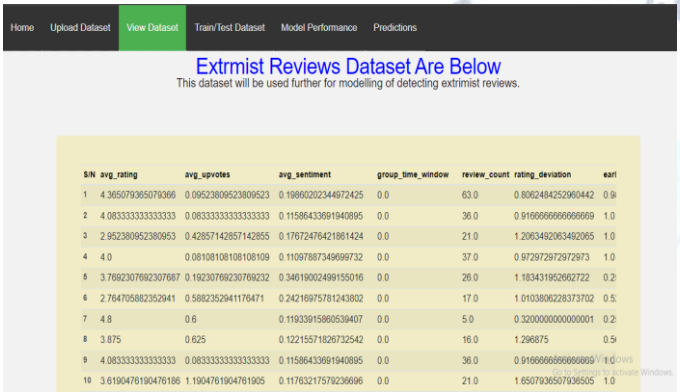
HOME PAGE



UPLOAD DATASET

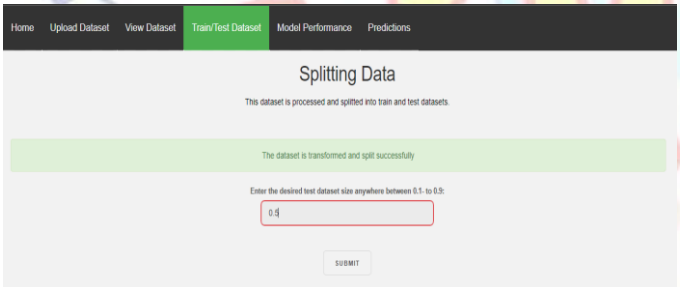


VIEW DATASET

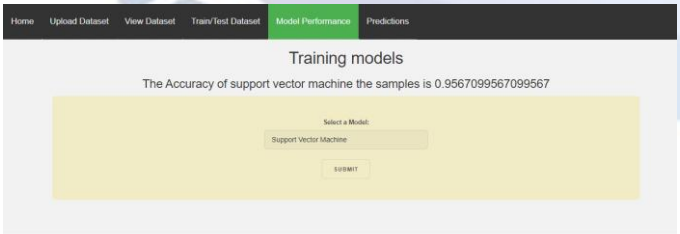


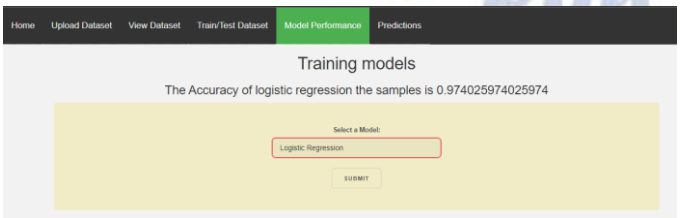
S/N	avg_rating	avg_upvotes	avg_sentiment	group_time_window	review_count	rating_deviation	earl
1	4.365079365079366	0.095238095238095 23	0.19860202344972425	0.0	63.0	0.8062484252960442	0.9
2	4.083333333333333	0.08333333333333333	0.11586433691940895	0.0	36.0	0.9166666666666669	1.0
3	2.952380952380953	0.42857142857142855	0.17672476421861424	0.0	21.0	1.2063492063492065	1.0
4	4.0	0.08108108108108109	0.11097887349699732	0.0	37.0	0.972972972972973	1.0
5	3.7692307692307687	0.1923076923076923 2	0.3461900249915501 6	0.0	26.0	1.183431952662722	0.2
6	2.764705882352941	0.5882352941176471	0.24216975781243802	0.0	17.0	1.0103806228373702	0.5
7	4.8	0.6	0.11933915860539407	0.0	5.0	0.3200000000000001	0.2
8	3.875	0.625	0.12215571826732542	0.0	16.0	1.296875	0.5
9	4.083333333333333	0.08333333333333333	0.11586433691940895	0.0	36.0	0.9166666666666669	1.0
10	3.619047619047618 6	1.1904761904761905	0.11763217579236696	0.0	21.0	1.6507936507936505	1.0

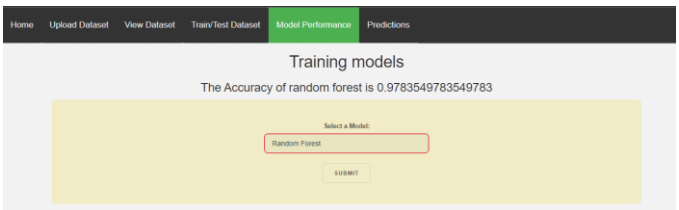
TRAIN/TEST DATASET

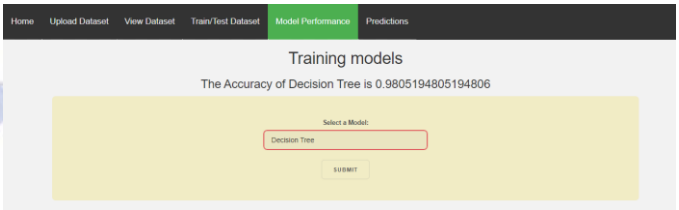


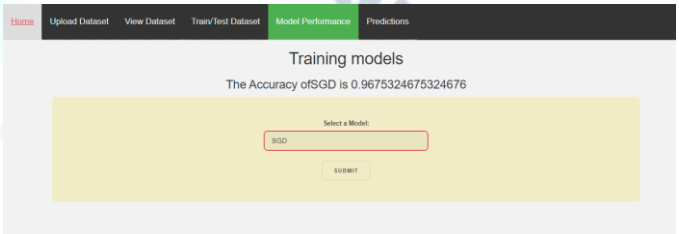
MODEL PERFORMANCE

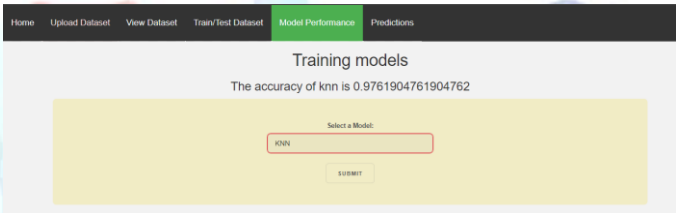


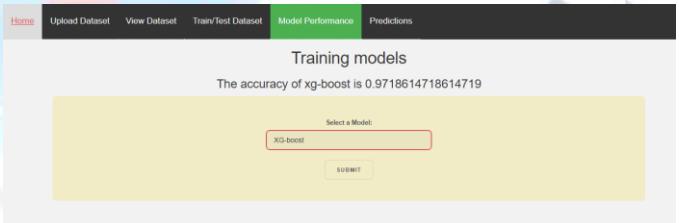




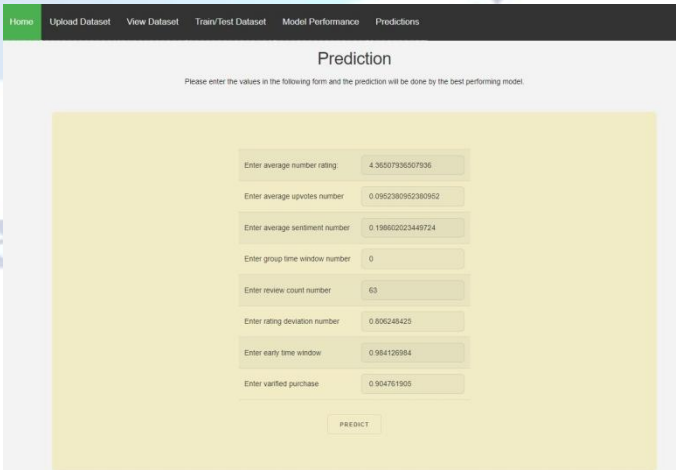


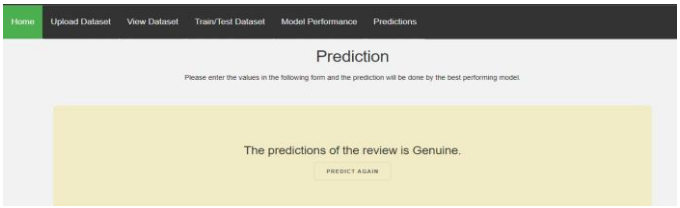


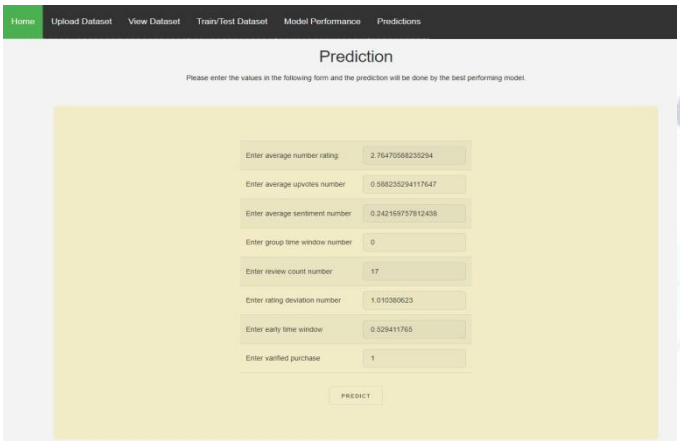


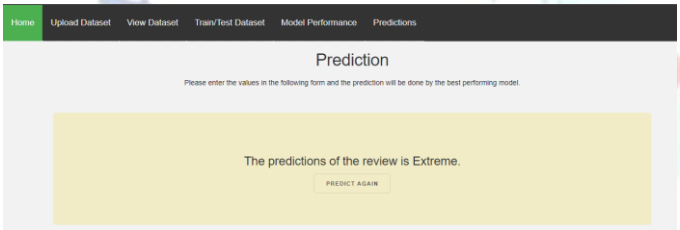


PREDICTIONS









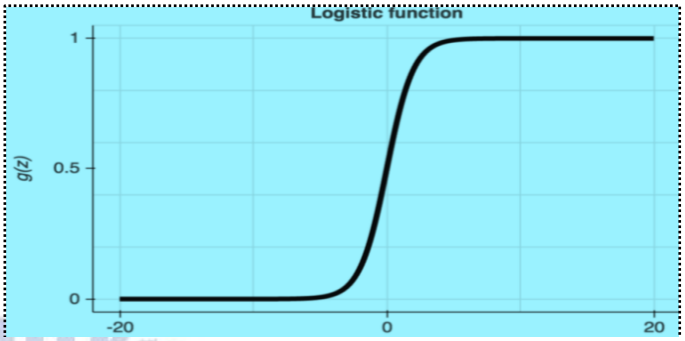
4.1. Algorithms

4.1.1. Analysis of logistic regression

When dealing with categorical dependent variables, the chance of occurrence is predicted using a technique called as logistic regression. The dependent variable in a logistic regression is binary data, and the dependent variable is binary data (no, failure, etc.). The logistic regression model predicts that P(Y=1) will be equal to one in the presence of the independent variable X.

STEP1: HYPOTHESIS OF THE LOGISTIC REGRESSION ANALYSIS

The logistic regression classifier may be derived by comparing it to the function g of the logistic regression (z). As the y-axis approaches the asymptotes of the logistic function, the value 0.5 marks the asymptotic point [22].



1. **Figure.4.1. Graph of LR function**

Step (1b): Decision framework for logistic regression

In our data set, we have two characteristics: height and weight, according to the logistic regression hypothesis.

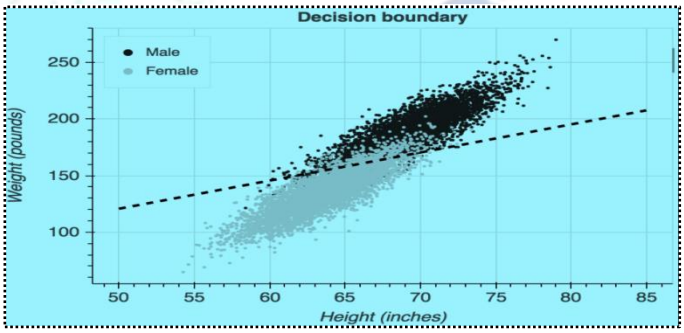


Figure.4.2. Graph of Decision framework for logistic regression

4.1.2. Classifier based on a random forest

It is possible to attain outstanding results in both classification and regression problems when employing the random forest method. As opposed to this, the most usual situation in which this method is employed is when there is a classification issue [16]. We are all known that the bigger the number of trees in a forest, the greater the strength of the forest. An individual decision tree for each data sample is generated using the random forest algorithm, which then votes on which forecast is the most correct among all of the predictions. Due to over fitting reduction, the ensemble method outperforms a single decision tree in terms of over fitting reduction [19].

Working principle of Random Forest Algorithm

The stages that follow will assist us in gaining a better understanding of how the Random Forest algorithm works:

Step.1: A random sample can be selected from a dataset in order to get things started.

Step.2: Following that, using this approach, a decision tree will be constructed for each sample. After that, the prediction results for each decision tree will be retrieved.

Step.3: At the conclusion of this stage, there will be a vote for each anticipated occurrence that has been forecasted.

Step.4: Finally, select the prediction result that obtained the greatest number of votes.

4.1.3. SGD

Prior to discussing Stochastic Gradient Descent, it is necessary to define the term Gradient Descent (GD). Gradient Descent, an optimization strategy for learning algorithms that has been applied in a variety of circumstances, can be beneficial to a wide range of algorithms and situations. It is the slope of a function that determines the gradient of a function. With the help of another variable, it is feasible to determine how much a variable changes in reaction to a change in another. Gradient Descent returns the partial derivative of the parameters that are passed to it as input because it is a convex function.

4.1.4. XG-BOOST

In the current state of affairs, the machine learning algorithm XGBoost is the most often used. No matter what type of data they're analyzing, it's commonly acknowledged that it outperforms other machine learning algorithms. Structured data has been the "true love" of Kaggle users since the sites inception in early 2014.

Extensive Gradient Boosting (Extreme GBoost) is a distributed gradient boosting library that is highly efficient in terms of performance. Every aspect of the programme is guided by the GBM framework. The GBM framework, on the other hand, performs better on its own than it does on any other framework. XGBoost was created by Tianqi Chen, a doctoral candidate at the University of Washington

how much progress the algorithm makes per cycle, is critical for the algorithm's performance.

4.1.5. K-NEAREST NEIGHBOUR

Parts 1 and 2 of this k-Nearest Neighbours tutorial are divided into three sections:

Step.1: Determine the Euclidean distance.

The first step is to calculate the distance between two sets of information. Using a straight line between two data rows or vectors is the quickest and most straightforward technique to find out how far apart they are from one another. Even at a higher dimensional level, this makes intuitive sense and is straightforward to put into practice.

$$\text{Euclidean Distance} = \text{sqrt (sum i to N } (x1\_i - x2\_i)^2)$$

Step.2: Find Your Nearest Neighbors.

We have identified the k occurrences of data that are the closest to it, as determined by our distance metric, as the data's new neighbors. It is necessary to first estimate the distance between each new piece of data and the other items in our collection before we can discover the neighbors of that new piece of data.

Step.3: Make Forecasts

Predictions can be produced based on the data points in the training dataset that are the most similar to each other. Because of our search, we are able to offer you with the classification that is the most commonly used among the neighboring classifications. This can be accomplished by using the max() function to a list of output values that are close to each other. Make use of the max() function to obtain a list of class values.

5. CONCLUSIONS

During our research, we discovered a new type of opinion spam where spammers create negative reviews of entire businesses to influence public perception. These organisations are likely to have a large impact on many businesses' popularity and reputation on review websites. We've learned a lot about customer behaviour since discovering the link between brand-level group activity and scathing reviews. With this information readily available, internet reviews may now be used to make more accurate choices. This study found that selecting credible candidates using FIM and then analysing their collective behaviour could detect extreme spam organisations.

This enables us to train an algorithm to identify extremist organisations using human annotated labels as a ground truth. We then examined the accuracy of the classification algorithms for the severe and moderate categories using a number of categorization approaches. This issue, as well as the general elements of how these organisations assault the brands they want to target, were thoroughly investigated. We have decided to make the code and annotated data set available to future scholars.

5.1. Future scopes

Most extreme groups tend to be more concerned with their own image than the reputations of their adversaries. It would take a tremendous amount of time and resources to have an impact on a huge number of brands. Rather than competing with others, they found that improving their own brand image was more profitable and successful. Some of the future scopes is following below:

1. In order to obtain a better grasp on the situation, researchers dug deep into this incident and the broader patterns of how these groups attack these businesses.
2. It is necessary to conduct further research into the behaviour of these organisations because of the complexity of brand-level opinion fraud.
3. Consistency in ratings, sentiment in reviews, confirmed purchases, review dates, and helpful votes collected on reviews are all examples of these behaviours. We're starting to question if Amazon's present systems for reducing unofficial incentives may be overcome because so many certified reviewers voice strong opinions.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

References

[1] A. Kim. (2017). That review you wrote on Amazon? Priceless. [Online]. Available: https://www.usatoday.com/story/tech/news/2017/03/20/review-you-wrote-amazon-pricess/99332602/

[2] E. Gilbert and K. Karahalios, "Understanding deja reviewers," in Proc. ACM Conf. Comput. supported Cooperat. Work (CSCW), 2010, pp. 225–228, doi: 10.1145/1718918.1718961.

[3] Amazon.in. (2018), Review Community Guidelines. [Online]. Available: https://www.amazon.in/gp/help/customer/display.html?nodeId=201929730

[4] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in Proc. 21st Int. Conf. World Wide Web (WWW), 2012, pp. 191–200.

[5] Y. Lu, L. Zhang, Y. Xiao, and Y. Li, "Simultaneously detecting fake reviews and review spammers using factor graph model," in Proc. 5th Annu. ACM Web Sci. Conf. (WebSci), 2013, pp. 225–233.

[6] S. Rayana and L. Akoglu, "Collective opinion spam detection: Bridging review networks and metadata," in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2015, pp. 985–994.

[7] S. Dhawan, S. C. R. Gangireddy, S. Kumar, and T. Chakraborty, "Spotting collective behaviour of online frauds in customer reviews," 2019, arXiv:1905.13649. [Online]. Available: http://arxiv.org/abs/1905. 13649

[8] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in Proc. 12th Int. Conf. World Wide Web (WWW), 2003, pp. 519–528.

[9] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proc. Conf. Empirical Methods Natural Lang. Process., 2002, pp. 79–86.

[10] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, "Sentiment analysis and classification based on textual reviews," in Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES), Feb. 2013, pp. 271–276.

[11] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," Expert Syst. Appl., vol. 36, no. 3, pp. 6527–6535, Apr. 2009.

[12] M. Chelliah and S. Sarkar, "Product recommendations enhanced with reviews," in Proc. 11th ACM Conf. Recommender Syst., Aug. 2017, pp. 398–399.

[13] L. Chen and F. Wang, "Preference-based clustering reviews for augmenting e-commerce recommendation," Knowl.-Based Syst., vol. 50, pp. 44–59, Sep. 2013.

[14] J. Feuerbach, B. Loepp, C.-M. Barbu, and J. Ziegler, "Enhancing an interactive recommendation system with review-based information filtering," in Proc. IntRS@RecSys, 2017, pp. 10–55.

[15] Pattanayak, R. M., Panigrahi, S., & Behera, H. S. (2020). High-order fuzzy time series forecasting by using membership values along with Data and Support Vector Machine. Arabian Journal for Science and Engineering, 45(12), 10311-10325.

[16] Pattanayak, R. M., Behera, H. S., & Panigrahi, S. (2021). A novel probabilistic intuitionistic fuzzy set based model for high order fuzzy time series forecasting. Engineering Applications of Artificial Intelligence, 99, 104136.

[17] Pattanayak, R. M., Behera, H. S., & Panigrahi, S. (2021). A Non-Probabilistic Neutrosophic Entropy-Based Method For High-Order Fuzzy Time-Series Forecasting. Arabian Journal for Science and Engineering, 1-23.

[18] S. Panigrahi, R.M. Pattanayak, P.K. Sethy, S.K. Behera. (2021). Forecasting of Sunspot Time Series Using a Hybridization of ARIMA, ETS and SVM Methods, Solar Physics, 296(1), 1-19.

[19] Pattanayak, R.M., Sangameswar, M.V., Vodnala, D., Das, H. (2021). Fuzzy Time Series Forecasting Approach using LSTM Model. Computacion y Sistemas.

[20] Pattanayak, R. M., Behera, H. S., & Panigrahi, S. (2020). A multi-step-ahead fuzzy time series forecasting by using hybrid chemical reaction optimization with pi-sigma higher-order neural network. Computational intelligence in pattern recognition, 1029-1041.

[21] Pattanayak, R. M., Behera, H. S., & Panigrahi, S. (2020). A novel hybrid differential evolution-PSNN for fuzzy time series forecasting. In Computational Intelligence in Data Mining (pp. 675-687). Springer, Singapore.

[22] Pattanayak, R. M., Behera, H. S., & Rath, R.K.: A Higher Order Neuro-Fuzzy Time Series Forecasting Model Based on Un-equal Length of Interval. In: International Conference on Application of Robotics in Industry using Advanced Mechanisms. pp. 34–45 (2020). SPRINGER International Publishing.

[23] Pattanayak R.M., Behera H.S.: Higher order neural network and its applications: A comprehensive survey. In: Advances in Intelligent Systems and Computing. pp. 695–709 (2018), Springer, Singapore. SPRINGER International Publishing.

[24] Parvathi, D. S. L., Leelavathi, N., Ravikumar, J. M. S. V., & Sujatha, B. (2020, July) Emotion Analysis Using Deep Learning. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 593-598). IEEE.

[25] Kumar, J. R., Sujatha, B., & Leelavathi, N. (2021, February). Automatic Vehicle Number Plate Recognition System Using Machine Learning. In IOP Conference Series: Materials Science and Engineering (Vol. 1074, No. 1, p. 012012). IOP Publishing."