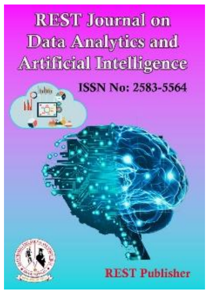
REST Journal on Data Analytics and Artificial Intelligence

Vol: 2(2), June 2023

REST Publisher; ISSN: 2583-5564

Website: https://restpublisher.com/journals/jdaai/

DOI: https://doi.org/10.46632/jdaai/2/2/15



Finding Harmful Comments on Social Networking Sites Using NLP and Machine Learning Methods

[1]Esha Bansal, *[2]Nisha Bansal

[1]*Government College for Women,Jind. Haryana, India.*

[1]*I.T.S Mohan Nagar Ghaziabad, Uttar Pradesh, India*

*Corresponding Author Email: nishabansal20@gmail.com

Abstract: *The usage of violent language has significantly increased due to social media and networking. A key component in this is the younger generation. More than half of young people who use social media are affected by cyberbullying. Harmful interactions occur as a result of insults expressed on social networking websites. These comments foster an unprofessional tone on the internet, which is usually understood and mitigated through passive mechanisms and techniques. Additionally, the recall rates of current systems that combine insult detection with machine learning and natural language processing are incredibly poor. To establish a viable classification scheme for such concepts, the research analyzes how to identify bullying in writing by examining and testing various approaches. We propose an effective method to assess bullying, identify aggressive comments, and analyze their veracity. NLP and machine learning are employed to examine social perception and identify the aggressive impact on individuals or groups. The ideal prototyping system for identifying cyber dangers in social media relies heavily on an efficient classifier. The goal of the paper is to emphasize the critical role that learning strategies play in enhancing natural language processing efficiency.*

1. INTRODUCTION

Vision for computers and natural language processing are only two areas of artificial intelligence that have seen significant advancements because to deep learning and machine learning. in addition to having a positive impact on numerous business sectors. These learning techniques are crucial for proper analysis since natural language processing, which is the capacity of computer systems to comprehend human languages, is a challenging undertaking. Computer science's Natural Language Processing (NLP) field focuses on automated text and language analysis. Due to ongoing advancements in deep learning and machine learning, NLP approaches have advanced significantly in recent years. The employment of reasonable Language Processing (NLP) techniques in social networking sites seems reasonable given these parallels in form and content. Even though the word "NLP" relates to natural languages, programming code is studied using the same computational techniques. Over the past two decades, The discipline of NLP has continuously contributed statistical and machine learning methods to bioinformatics. The application of machine learning and natural language processing (NLP) techniques has greatly facilitated the processing of digital data. Because of the growing reliance on digital data, it is essential to maximize its potential in numerous study domains. Social media platforms can be utilized for a variety of tasks, including information extraction from text, automatic vocabulary management, text extraction from social media platforms, data mining, research object identification, and analysis. And its consequences, etc. Even though NLP-based machine learning approaches perform very well in the biomedical and healthcare industries, additional expertise is needed to analyze narrative medical material [1]. In order to generate new opportunities in this field, it is vital to thoroughly examine the issues and difficulties with information extraction from medical literature [2]. The youthful population has escalated the usage of insults or cyberbullying on websites. Social media usage is becoming more and more prevalent every day. Both the number of users and the amount of time they devote to the site are growing tremendously. They are unrestricted and able to speak whatever they want on this platform. Access to this website or social platform is impacted for other users. Limiting the information on this social media platform that uses cyberbullying terminology through analysis will enable the constructive discourse for which this social media platform was designed to be used. The way that people share their opinions on issues has changed as a result of the abundance of public platforms that are available online. It made it simple to swiftly share our opinions with a sizable group regarding the general public or a specific person. But it's also important to take into account how quickly insensitive remarks and ideas can spread in a public setting. Cyberbullying was neglected and not taken seriously in the previous year. Due to the poor user activity, it was advised to screen off or disconnect if any threatening comments were made. But right now, everything is totally different. These comments are being used by 50% of social media users in 2017, and this cannot be disregarded. This major issue is also being experienced by popular social media platforms like Twitter and Facebook. These derogatory terms have been used against people and communities, and numerous legal suits have been brought.

2. SOCIAL NETWORKING PLATFORMS PROBLEM STATEMENT

Platforms for social networking have become an essential part of people's life since they give them a way to communicate, share information, and connect with others. These platforms do, however, also deal with a number of problems and difficulties that may have an effect on both its users and society at large. One of the main difficulties with social networking platforms is the spread of misleading information and fake news. False information and propaganda have thrived on social media, which can significantly affect public opinion and even influence elections and political results. Cyberbullying is another problem; when people use social media to harass, intimidate, and abuse others, it can have a bad impact on their mental health and even result in suicide in some circumstances. Another issue is privacy because social media sites frequently gather enormous amounts of user-provided personal data that may be sold to third parties or used for targeted advertising. Users may believe that their personal data is not sufficiently protected, which can cause problems with data privacy and security. With social media systems built to keep users interested and skimming through content for as long as possible, there is also the issue of addiction. As a result, there may be detrimental effects on mental health, such as depression, anxiety, and social isolation. While social networking sites have generally benefited society, In order to continue being a constructive influence in the years to come, they must also deal with a variety of problems. Because so many people utilize the internet, cyber security is becoming a bigger issue. We are aware that websites offer quick, convenient, interactive access to online social networking sites from everywhere, making them an appealing target for cybercrimes. Cyberbullying refers to material or visuals that are posted on socially unacceptable websites. Words that are hurtful to both a group and an individual are considered abusive. On websites like Twitter, YouTube, Instagram, and many more, people use antisocial language and present their opinions in an extremely nasty or combative way. This is yet another instance of online harassment. It is regarded as an internet danger in the US. Some of the biggest challenges in preventing cyberbullying include finding these words and phrases on internet platforms, presenting these cases to legal authorities, and identifying the perpetrators. On social media or online community platforms, there is currently no system in place to automatically and intelligently recognize instances of aggression and online harassment. This significant problem was not regarded as a research problem because of how bad it was, but it is now in a critical stage. There is no way to disregard its impact on the website. Researchers and cybercrime organizations must pay close attention to this conduct if it is to be stopped. The goal of the research project is to develop a prototype that can automatically detect cyberbullying and inappropriate conduct in social media and online groups, thereby preventing online harassment and violence. First, the data set must be extracted, gathered, and labeled. 2. Numerous characteristics are tested, cleaned, and prepped to improve accuracy. 3. Classifying text, remarks, or postings into different groups. 4. Assessment and analysis of the ideal model. The project's objective is to gain knowledge on utilizing natural language processing and machine learning to a relevant problem, such as online bullying and harassment.

3. BUILDING MACHINE LEARNING MODEL METHODOLOGY

We propose a novel method for message analysis and detection of potentially harmful activities that combines machine learning and natural language processing. The pipeline consists of extracting pertinent data from numerous internet sources, pre-processing, generating ground truth, engineering and selecting features, and classifying the data. The objective of this supervised learning issue is to divide online user content into "Bully" and "Non-Bully" categories for status and post updates. Collecting pertinent information from many websites online is a crucial first step. The majority of this sort of data is composed of user posts, comments, photos, videos, and audios. As was previously said, gathering raw data sets is the initial step in identifying online threats like hate speech and insults. The great majority of user comments, posts, images, and videos on social networking and other social media sites provide data sets for cyberbullying. Two resources for discovering a large range of data sets are Kaggle, both the UCI Machine Learning Repository, which includes millions of open source data sets for data analysis applications, and the places where individuals and companies submit data for research and competition.

Building Machine Learning Model: The classification results for the training dataset along with the test dataset that Imperium obtained on Kaggle are displayed in Table 1. The table shows the various performance evaluation metrics that were used after training and testing the datasets. Although the test accuracy of all four classifiers is

between 50% and 55%, their training accuracy ranges from 75% to 90%. Put simply, data preparation is the process of transforming raw data into a format that can be understood. Real-world data are often inaccurate because they lack specific actions or trends. inconsistent, and/or all of these things. Preprocessing data has been successfully used to address these issues. Raw data is prepared for further processing through data preparation.
Count vectors and TF-IDF vectors are used to combine all of the retrieved feature vector sets into a single feature set with up to five levels of n-gram ordering as tokens of two words and letters. The training dataset has a total of (6594, 4600) features or predictors, where 6594 is the number of samples and 4600 is the total number of features. There are also 2235, 4600 elements in the test dataset. Logistic regression, Support Vector Machine, and two well-liked ensemble methods, Random Forest Classifier and Gradient Boosting Machine, are the classifiers with the most basic classifications. for the goal of creating a machine learning model. Random forest and gradient boosting machines need dense feature set matrices, but logistic regression and support vector machines need sparse feature set matrices. In order to create a dense matrix for training, the space matrix of feature vectors is correctly converted. Numerous higher criteria are researched and altered in order to increase learning effectiveness. For instance, the regularization strength is the reverse of "C," a parameter used in logistic regression and support vector machines. Stronger regularization is indicated by smaller values in SVM. Other variables in a gradient boosting machine include "the number of trees in a forest," "the learning rate," "the number of subsamples," etc. The models are created using the Python versions of the aforementioned classifiers from Scikit-learn. Examples of advertising categories. On the basis of the training data, a support vector machine is trained. Given that the training dataset is unbalanced and comprises mainly of negative examples, the cost factor determines how much the cost of an error on a positive example should be greater than an error on a negative example. Parameter tuning using the hit-and-miss method, with "C" values of [0.002, 0.02, 0.003, 0.03, 300] and "J," the cost factor, of [10, 30, 100], being used as assumptions. For all four models, the training period differs considerably. SVM required 6.668 seconds for training, compared to 0.060 seconds for logistic regression. The two ensemble approaches, random forest and gradient boosting machines, took the longest, 196.236 seconds and 324.447 seconds, respectively. After being trained, the models were used with the Kaggle test dataset. The predictions generated for the test dataset are then saved in a file. A score of 0 to 0.5 indicates a "non-insulting" opinion, whereas a score of 0.5 to 1 indicates a "insulting" opinion. All predictor values lie between 0 and 1.

4. NATURAL LANGUAGE PROCESSING ALGORITHMS

Following the pre-processing of your data, you should create an NLP algorithm and train it to understand natural language and carry out certain tasks.
To handle NLP issues, you can employ two primary algorithms:

1. A rule-based strategy is: Grammar rules that are manually written by linguists or knowledge engineers are the foundation of rule-based systems. This method of developing NLP algorithms was the first and is still in use today.
2. Algorithms for machine learning: Machine learning models, on the other hand, are based on statistical methods and learn to complete tasks after being fed examples (training data).

The primary advantage of machine learning algorithms is without a doubt their capacity for self-learning. You have more freedom because there is no need to establish explicit rules because machines can make predictions on their own using historical data. Machine learning algorithms are trained to create connections between a specific input and its related output by feeding them training data and expected results (tags). To produce predictions for yet-to-be-observed data (new texts), machines first determine which properties best characterize the texts, building their own "knowledge bank" in the process:

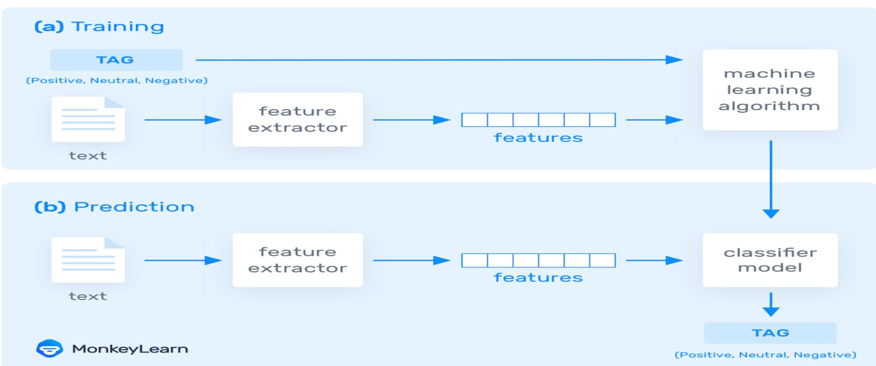


Figure 1. Building machine learning model

5. TOPSIS METHOD

In many sectors, Multi-Criteria Decision Making (MCTM) is quick. An expanding issue among the places. How to assess a set of alternatives based on several criteria is the main issue. Despite the fact that this issue is highly practical, there are some solutions, and it might be challenging to assess their quality. Multiple Targets are taken into account simultaneously for optimization in multi-criteria decision-making models. Each objective's measure is for other objectives. Size can change. For instance, increasing profitability, which is gauged in terms of money, could be a goal. Reducing personnel time could also be a goal. However, not all of these goals are in the same direction, and occasionally they conflict with one another. Better decision-making in this area comes from better programming. Due to its significance and practical use, an explanation of multi-criteria decision-making models is provided here. The most suitable of based on n criterion for decision available options other methods of selection or prioritization, and decision-making processes that consider multiple criteria and are common to the group. There are several multi-criteria strategies to help with selection in circumstances when there are multiple criteria. The abbreviation TOPSIS, which stands for Optimal Solution, denotes a prioritization method. Originally presented by Hwang and Yoon, TOPSIS is appealing to decision-makers because minimal subjective input is needed. Weights subject to judgment All that is needed is input. The three types of this technique attributes or standards that comprise: Benefit attributes that are qualitative and quantitative cost factors or two synthetic substitutes are taken into consideration in this procedure. Best Alternative: For each attribute that is taken into account, there is an ideal circumstance. Negative The worst possible option is one with subpar qualities. TOPSIS is focused on maximizing the distance from a nadir point while also shortening the distance from a great point. MCDM is a technique for discovering solutions from finite choices. TOPSIS is crucially important. Combining relative weights is possible. Imagine this time that we have n traits or criteria and m alternatives (options), and that each criterion has a score for each desire. Using the aforementioned concepts, in this way, starting from the ideal place an alternative's distance is considered, as well as its distance from the opposite of the ideal location. The distance between the chosen alternative's ideal solution and the negative solution must be very small and very large, respectively.

TABLE 1. Accuracy score of train and test dataset.

Model	Train Accuracy	Test Accuracy	AUC Score	Cross Validation
Logistic Regression	0.900	0.537	0.577	0.659
SVM	0.766	0.523	0.578	0.620
Random Forest	0.905	0.545	0.579	0.653
Gradient Boost	0.774	0.532	0.537	0.647

Table 1 GIVES the Data set of the alternatives is Logistic Regression, SVM, Random Forest, and Gradient Boost Evaluation Parameters of the Train Accuracy, Test Accuracy, AUC Score and Cross Validation.

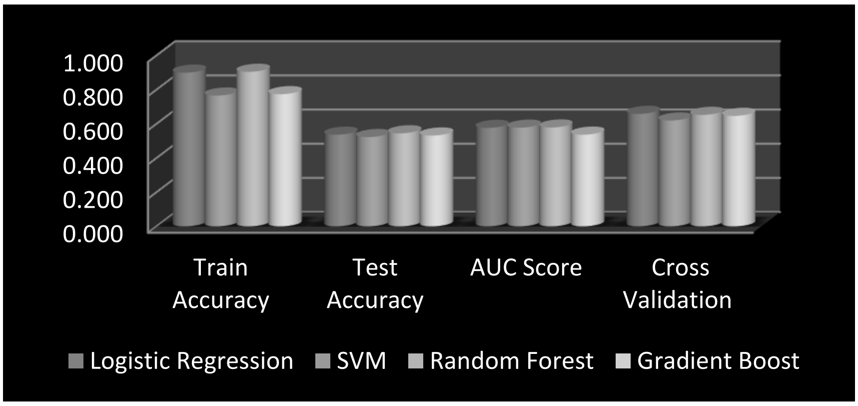


FIGURE 2. Data set graph

“Figure 1 shows that the Data set of the alternatives is Logistic Regression, SVM, Random Forest, and Gradient Boost Evaluation Parameters of the Train Accuracy, Test Accuracy, AUC Score and Cross Validation”.

TABLE 2. Normalized Data

	Train Accuracy	Test Accuracy	AUC Score	Cross Validation
Logistic Regression	0.5364	0.5025	0.5079	0.3928
SVM	0.4566	0.4894	0.5088	0.3695
Random Forest	0.5394	0.5100	0.5097	0.3892
Gradient Boost	0.4613	0.4978	0.4727	0.3856

"Table 4 shows the data from which the normalized data is calculated from the data set value is divided by the sum of the square root of the column value".

TABLE 3. Weight

Logistic Regression	0.25	0.25	0.25	0.25
SVM	0.25	0.25	0.25	0.25
Random Forest	0.25	0.25	0.25	0.25
Gradient Boost	0.25	0.25	0.25	0.25

Table 3 shows the weight of the data set the weight is equal for all the value in the set of data in the table 1. The weight is multiplied with the previous table to get the next value.

TABLE 4. Weighted normalized decision matrix

	Train Accuracy	Test Accuracy	AUC Score	Cross Validation
Logistic Regression	0.1341	0.1256	0.1270	0.0982
SVM	0.1141	0.1224	0.1272	0.0924
Random Forest	0.1349	0.1275	0.1274	0.0973
Gradient Boost	0.1153	0.1245	0.1182	0.0964

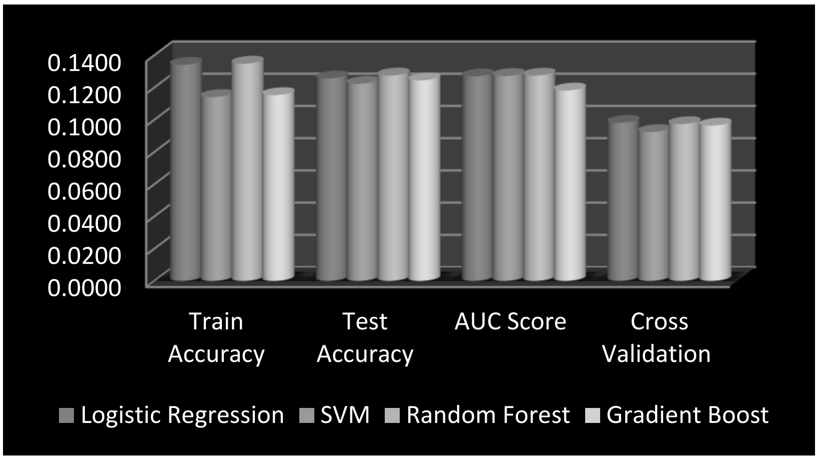


FIGURE 3. Weighted normalized decision matrix

TABLE 5. Positive Matrix

	Train Accuracy	Test Accuracy	AUC Score	Cross Validation
Logistic Regression	0.1349	0.1275	0.1274	0.0982
SVM	0.1349	0.1275	0.1274	0.0982
Random Forest	0.1349	0.1275	0.1274	0.0982
Gradient Boost	0.1349	0.1275	0.1274	0.0982

Table 5 shows the positive matrix of the data set that is calculated from the weighted normalized result matrix by calculating the maximum and minimum of the benefit factor and the cost factor.

TABLE 6. Negative matrix

	Train Accuracy	Test Accuracy	AUC Score	Cross Validation
Logistic Regression	0.1141	0.1224	0.1182	0.0924
SVM	0.1141	0.1224	0.1182	0.0924
Random Forest	0.1141	0.1224	0.1182	0.0924
Gradient Boost	0.1141	0.1224	0.1182	0.0924

Table 6 shows the negative matrix of the data set that is calculated from the weighted normalized result matrix by calculating the minimum and maximum of the benefit factor and the cost factor.

TABLE 7. SI plus, SI Negative, and Ci

	SI Plus	Si Negative	Ci
Logistic Regression	0.0021	0.0228	0.9171
SVM	0.0221	0.0090	0.2897
Random Forest	0.0009	0.0238	0.9638
Gradient Boost	0.0219	0.0047	0.1766

Table 7 show the sum of the calculation positive and negative matrix, the Si plus is calculated from the positive matrix, Si negative is calculated from the negative matrix and the Ci is calculated from the sum of the Si plus and Si negative.

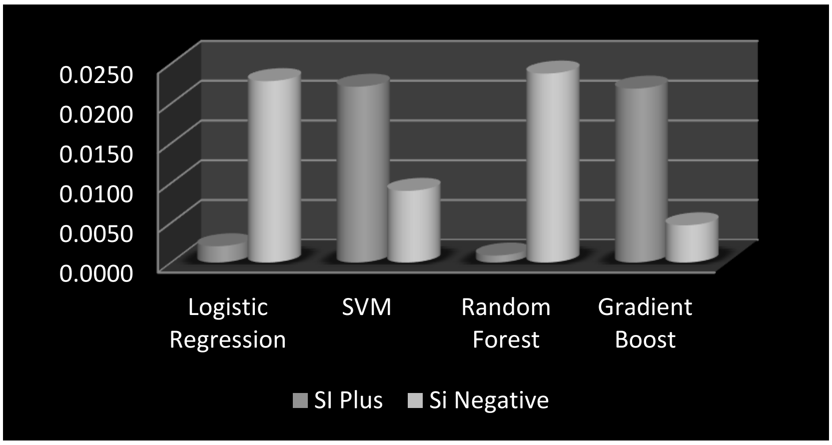


FIGURE 4. SI plus and SI Negative

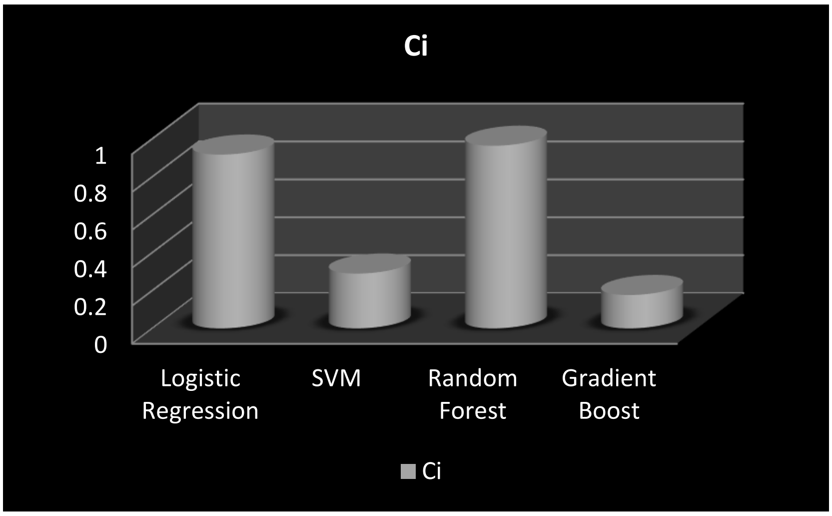


FIGURE 5. Ci Values

Figure 4 and 5 show the sum of the calculation positive and negative matrix, the Si plus is calculated from the positive matrix, Si negative is calculated from the negative matrix and the Ci is calculated from the sum of the Si plus and Si negative.

TABLE 8. Rank

	Rank
Logistic Regression	2
SVM	3
Random Forest	1
Gradient Boost	4

Table 8 given the Random Forest is on 1st rank, Logistic Regression is on 2rd rank, SVM is on 3nd rank, and Gradient Boost is on 4th rank.

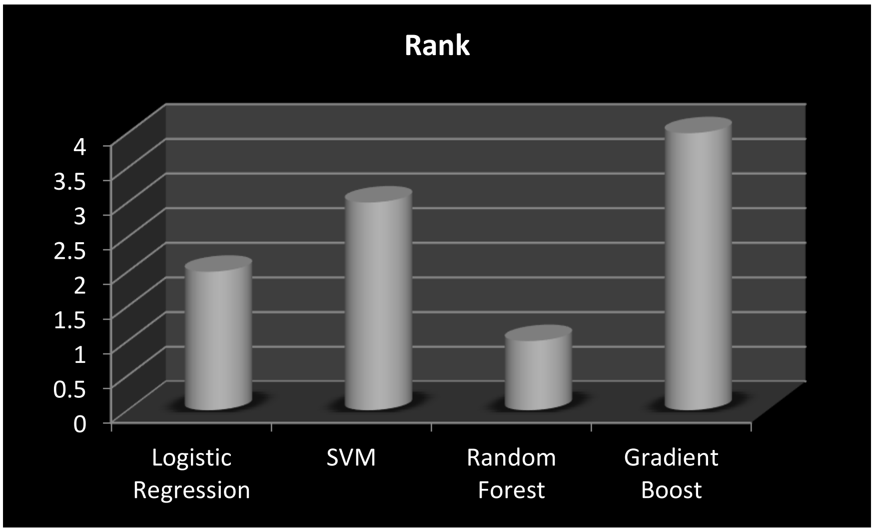


FIGURE 6. Ranking

Figure 5 given the Random Forest is on 1st rank, Logistic Regression is on 2rd rank, SVM is on 3nd rank, and Gradient Boost is on 4th rank.

6. CONCLUSION

One of the most exciting areas of artificial intelligence is natural language processing, which is already used in many daily applications like chat bots and search engines. With the help of NLP, companies may streamline some routine tasks and make the most of their unstructured data, collecting valuable insights that can be applied to raise client happiness and enhance customer experiences. Despite being a hard science, NLP is becoming more and more approachable for consumers because to online tools like MonkeyLearn, which make it simpler to create personalized models for tasks like text extraction and classification. Among the examined methods, experiments and overall project work show that Logistic Regression and Random Forest Classifier trained on feature layer outperform Support Vector Machine and Gradient Boosting Machine in this particular instance.

REFERENCES

[1]. Nagarhalli, Tatwadarshi P., Vinod Vaze, and N. K. Rana. "Impact of machine learning in natural language processing: A review." In *2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV)*, pp. 1529-1534. IEEE, 2021.

[2]. Ofer, Dan, Nadav Brandes, and Michal Linial. "The language of proteins: NLP, machine learning & protein sequences." *Computational and Structural Biotechnology Journal* 19 (2021): 1750-1758.

[3]. Garg, Ravi, Elissa Oh, Andrew Naidech, Konrad Kording, and Shyam Prabhakaran. "Automating ischemic stroke subtype classification using machine learning and natural language processing." *Journal of Stroke and Cerebrovascular Diseases* 28, no. 7 (2019): 2045-2051.

[4]. Goldberg, Simon B., Nikolaos Flemotomos, Victor R. Martinez, Michael J. Tanana, Patty B. Kuo, Brian T. Pace, Jennifer L. Villatte et al. "Machine learning and natural language processing in psychotherapy research: Alliance as example use case." *Journal of counseling psychology* 67, no. 4 (2020): 438.

[5]. Houssein, Essam H., Rehab E. Mohamed, and Abdelmgeid A. Ali. "Machine learning techniques for biomedical natural language processing: a comprehensive review." *IEEE Access* 9 (2021): 140628-140653.

[6]. Hodorog, Andrei, Ioan Petri, and Yacine Rezgui. "Machine learning and Natural Language Processing of social media data for event detection in smart cities." *Sustainable Cities and Society* 85 (2022): 104026.
[7]. Aone, Chinatsu, Mary Ellen Okurowski, and James Gorlinsky. "Trainable, scalable summarization using robust NLP and machine learning." In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pp. 62-66. 1998.
[8]. Ebadi, Ashkan, Pengcheng Xi, Stéphane Tremblay, Bruce Spencer, Raman Pall, and Alexander Wong. "Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing." *Scientometrics* 126 (2021): 725-739.
[9]. Olthof, Allard W., Prajakta Shouche, Eelco M. Fennema, Frank FA IJpma, RH Christian Koolstra, Vincent MA Stirler, Peter MA van Ooijen, and Ludo J. Cornelissen. "Machine learning based natural language processing of radiology reports in orthopaedic trauma." *Computer Methods and Programs in Biomedicine* 208 (2021): 106304.
[10]. Sharma, Hitesh Kumar, and K. Kshitiz. "Nlp and machine learning techniques for detecting insulting comments on social networking platforms." In *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pp. 265-272. IEEE, 2018.
[11]. Nikiforos, Stefanos, Spyros Tzanavaris, and Katia-Lida Kermanidis. "Virtual learning communities (VLCs) rethinking: influence on behavior modification—bullying detection through machine learning and natural language processing." *Journal of Computers in Education* 7 (2020): 531-551.
[12]. Bao, Yujia, Zhengyi Deng, Yan Wang, Heeyoon Kim, Victor Diego Armengol, Francisco Acevedo, Nofal Ouardaoui et al. "Using machine learning and natural language processing to review and classify the medical literature on cancer susceptibility genes." *JCO Clinical Cancer Informatics* 1 (2019): 1-9.
[13]. François, Thomas, and Eleni Miltsakaki. "Do NLP and machine learning improve traditional readability formulas?." In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pp. 49-57. 2012.
[14]. Ren, Lifeng, Yanqiong Zhang, Yiren Wang, and Zhenqiu Sun. "Comparative analysis of a novel M-TOPSIS method and TOPSIS." *Applied Mathematics Research eXpress* 2007 (2007).
[15]. Çelikbilek, Yakup, and Fatih Tüysüz. "An in-depth review of theory of the TOPSIS method: An experimental analysis." *Journal of Management Analytics* 7, no. 2 (2020): 281-300.
[16]. Zavadskas, Edmundas Kazimieras, Abbas Mardani, Zenonas Turskis, Ahmad Jusoh, and Khalil MD Nor. "Development of TOPSIS method to solve complicated decision-making problems—An overview on developments from 2000 to 2015." *International Journal of Information Technology & Decision Making* 15, no. 03 (2016): 645-682.
[17]. Jahanshahloo, Gholam Reza, F. Hosseinzadeh Lotfi, and Mohammad Izadikhah. "Extension of the TOPSIS method for decision-making problems with fuzzy data." *Applied mathematics and computation* 181, no. 2 (2006): 1544-1551.
[18]. García-Cascales, M. Socorro, and M. Teresa Lamata. "On rank reversal and TOPSIS method." *Mathematical and computer modelling* 56, no. 5-6 (2012): 123-132.
[19]. Chu, T-C., and Y-C. Lin. "A fuzzy TOPSIS method for robot selection." *The International Journal of Advanced Manufacturing Technology* 21 (2003): 284-290.
[20]. Chen, Pengyu. "Effects of normalization on the entropy-based TOPSIS method." *Expert Systems with Applications* 136 (2019): 33-41.

