**A Machine Learning Approach and Exploratory Analysis of Emotion (Depression & Sarcasm) using R Language.**

Arman Ansari1, a), Radhika Nambiar2, b), Vardaan Upadhyay1, c), Muskaan1, d), Chandan Raj Tripathi1, e), Chavali Sai Ram1, f)

Author Affiliations

1 Student, Dept. of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab - 144411

*2 Assistant Professor,* *Dept. of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab - 144411*

Author Emails

a) [armanansari21752@gmail.com](mailto:armanansari21752@gmail.com)

b) [radhika.28302@lpu.co.in](mailto:radhika.28302@lpu.co.in)   
c)[vardaanu7@gmail.com](mailto:vardaanu7@gmail.com)

d) [ranamuskaan2001@gmail.com](mailto:ranamuskaan2001@gmail.com)

e) [rajchandan452811@gmail.com](mailto:rajchandan452811@gmail.com)

f) [sairamchavali163@gmail.com](mailto:sairamchavali163@gmail.com)

**Abstract.** The study was conducted to analyse sentiment like depression and sarcasm to better comprehend and measure the strength of the communicated sentiments, whether it be positive, negative, or neutral. Finding and comprehending sarcasm in the text is the aim of sarcasm analysis. This research seeks to investigate several supervised machine learning algorithms to determine their significance and understand the many aspects of depression and sarcastic emotions amongst people which further can be analysed to know what the students nowadays are going through. This research will aid in the analysis, detection, prevention, and treatment of student sentiments. We reviewed over 20 papers and discovered that there are categories for the emotions that students express. What he or she is feeling is determined by the situation and his or her surroundings. The survey revealed that even though many students may not want to admit it, they are in the same situation as the others.

**Keywords**: Machine Learning, Student dataset, Depression, Sarcasm, Sentiment Analysis, Support Vector Machine, Random Forest, Decision Tree, Naïve Bayes, KNN, Supervised Machine Learning, R language, NLP (Natural Language Processing).

**INTRODUCTION**

Analysing digital text for sentimental analysis r12equires determining whether the message’s emotional tone is positive, negative, or neutral. Companies and institutions now have a lot of text data, such as emails, chat transcripts from customer service, students, comments on social media, and reviews[1]. This text can be scanned by sentiment analysis software to discover the author’s viewpoint on a subject. Businesses use the data from sentiment analysis to boost brand recognition, customer service and teacher student relationship. The volume of data delivered using all available resources is colossal. In this research we have narrowed down our study and focused on only Depression and Sarcasm, which will help us realize what emotions and sentiments are experienced by the youths of this generation[2]. We can observe after going through the statistics of WHO (World Health Organization), around 3.8% of the global populace is experiencing depressive issues among which is 5% of adults (4% among men and 6% among women), and 5.7% adults aged more than 60[3]. According to WHO around 280 million people in the world are going through depression but in the recent research the statistics shows that globally, 450 million people experience mental health issues, which includes depression as its wingman. According to a recent WHO survey, 4.1% and 4.4%, respectively, of Bangladesh’s population suffer from anxiety and 6.9 million from depression, or roughly 6.4 million and 6.9 million persons, respectively[4]. Sentimental analysis has become the most crucial area in the computer learning and processing language naturally [5]. The objective of sentimental analysis is to gather subjective and quantitative information from a variety of users, in this case specifically students, about their viewpoints, ideas for posts or other content, feedback forms, discussions, and conversations in order to determine what can be done to improve their situation and prevent future problems. Using figurative language on social media, particularly when using irony and sarcasm in clever and creative ways, poses substantial obstacles to these strategies. There have been many attempts to automatically detect the figure of speech, text, or conversation to help improve the overall sentiment analysis accuracy[6]. Sarcasm analysis is the process of identifying and understanding sarcastic language. Using machine learning and computational linguistics techniques, sarcasm is regularly identified and analysed in written or spoken language[7]. This is frequently done by examining several linguistic hints, such as lexical, syntactic, and pragmatic features. Since sarcasm frequently employs ambiguous or evasive language and can be highly context-dependent, analysis of it can be challenging. Based on a variety of factors, such as behavioural patterns, emotional states, and social contacts, depression analysis uses machine learning algorithms and techniques to evaluate and anticipate depression in individuals. To develop predictive models that can identify persons who are depressed or who may already be depressed, depression analysis is done[2]. This can help doctors intervene more quickly, provide patients with the treatment they need, and ultimately improve patient outcomes.

**LITERATURE REVIEW**

Sentiment analysis further narrowed down to depression and sarcasm analysis are both natural language processing techniques used to analyse text data, but they have different goals, advantages, and disadvantages.

The technique of identifying the emotional tone or attitude indicated in a text, such as a social media post or a customer review, is known as sentiment analysis. Finding the text's sentiment—whether it is a favourable negative, or neutral is the aim of sentiment analysis. This can be helpful for businesses to comprehend customer feedback on their goods or for social media monitoring to follow public opinion on a certain subject.

Irfan Ali Kandhro and his other group members suggested employing different machine learning models to analyse valuable data, and they also incorporated MNB and MLP in their investigation into text classification. Their research aids in raising the standard of instruction in the educational system. [8]

The outcome offered preliminary proof that training data for automated machine learning models for sarcasm, irony, and sentiment analysis were appropriate. This study used an SVM model with a 62% accuracy rate. [9]

All the previous research depicts few algorithms and their comparison on error rate and not on the precision of the model. Furthermore, there is no learning outcomes provided by the research for them to overcome the situation. The data is collected anonymously from the students so that their privacy is maintained, and we get a valid and viable data for the research purpose and provide a learning outcome accordingly.

**CONVENTIONAL TECHNIQUES FOR MACHINE LEARNING**

The use of natural language in interactions between people and machines is the focus of the artificial intelligence discipline of NLP (Natural Language Processing). In language processing, NLP helps eliminate ambiguity and gives data structured form. There are several conventional methods for analysing the raw data, but KNN, SVM, Naive Bayes, Decision Tree, and Random Forest are the ones we found to be most effective. Each of these algorithms has a unique method for predicting data, and they are all highly supervised.

Diagram

Description automatically generated

Figure : Flowchart representing the anticipated implementation work.

**PROPOSED TECHNIQUE.**

The task is carried out keeping in mind the workflow diagram[10] that we came up with, to orderly collect and analyse the dataset. The Data flow diagram (Figure 1) is briefly explained with the steps below:

Data tracking

The research's goal was defined as the initial phase to be carried out, we came up with the idea of collecting a real-world data of students from cross domain and distribute the google form amongst them.[11]

The data collected was distributed amongst the students of cross domain at Lovely Professional University. Around 110 records were collected keeping in mind that the more the data the more the noise will be[12], [13]. After the collection of sufficient records, the dataset was firstly exported to MS Excel and analysed their first-hand. Few pivot tables and charts were created to have a clear picture to what is to be done while analysing the same in R Studio.

Data Pre-processing:

* The first cleaning was done with the help of MS Excel by removing unwanted columns and sentiments that weren’t necessary for the analysis.
* The next step is to perform EDA by loading the dataset into R Studio for further pre-processing.
* After loading the data and required packages, we carry out important tasks including eliminating null values and text mining to prune or clean out any unnecessary information that doesn't signal the start of the procedure.

Factorization:

For better evaluation of the model some attributes are converted into factor of -1, 0 and 1. The column named Diagnosis contained “Yes” and “No” which then was converted into factor of level 1 and 0 respectively. Similarly, the attribute named Affect contained values “Negatively”, “No Impact” and “Positively” were converted into factors of level -1, 0 and 1 respectively[14]. Furthermore, the columns “sarcasm\_ambiguity” and “recognize\_sarcasm” were also converted into factors.

Analysing the dataset

The dataset's structure, number of observations, dependant variables, and independent variables were evaluated using statistical techniques in R. To determine the degree of the features, research was conducted, and a few conclusions were drawn[15]. The data investigation's findings are explained in the Findings and Discussion section. After the Data pre-processing we split by dividing the dataset into train and test sets as we are using supervised machine learning algorithms where the test set serves to forecast the model's result based on the test set and the train set serves as the basis for train the model.

Model Building

A thorough analysis shows that each dataset parameter significantly affects the dependant variable for each condition. As a result, each variable is considered while creating a prediction model for each scenario. Before these models are formed, the sample is randomly split into testing and training batches. The suggested project was developed using the required R programming libraries and functions for numerous supervised machine-learning classification-based methods[16]. After building the model and predicting the values of the test data, we then evaluate the accuracies of each model and compare them on a bar plot.

The Algorithm used to evaluate or analyse the data are:

* KNN (K-Nearest Neighbour)
* Decision Tree
* Random Forest
* Naïve Bayes
* Support Vector Machine (SVM)

Model Selection:

The analysis is carried out by applying all the above models and predict the outcome and accuracies of each data. We then evaluate the accuracy of each model to determine which model delivers the best predictions and what results is produced after analysis. After comparing of each model, we concur that the most accurate model, Random Forest, boasts a precision of 94.54% of depression dataset and 72.72% of sarcasm dataset. This can be used as a steppingstone for the frontier area to advance with the research.

Summary of Work:

* Objective of the project
* Collecting the dataset though distributing google forms amongst students of cross domain.
* Preparing the data
* Examining the properties of data
* Evaluate the dataset using different machine learning algorithms.
* Model Selection
* Its application in practical world.

**RESULT**

There has been numerous research conducted on the topic of sentiments where the researchers have used various algorithms and found out different accuracies for their respective model. Similarly, we have also worked on real life data and analysed the best prediction model out of all, here is a brief summary table of the accuracies of respective model that we have used.

Table 1: Percentage accuracy of various algorithm

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No** | **Model** | **Depression Accuracy** | **Sarcasm Accuracy** |
| 1. | K-NN | 52.72727 | 1.818182 |
| 2. | SVM | 21.42857 | 28.57143 |
| 3. | Random Forest | 94.54545 | 72.72727 |
| 4. | Naïve bayes | 88.88889 | 55.55556 |
| 5. | Decision Tree | 8.696969 | 13.04348 |



Figure 2: Word cloud of NLP model

Figure 2 is word cloud that tells us about the frequency of words used in the reviews attribute that were provided by the students in the google form.

Chart, histogram

Description automatically generatedFigure 3: Frequency Bar plot of reviewed terms reflecting sentiments in NLP model.

Chart, bar chart

Description automatically generated

Figure 4: Graph of accuracy of various algorithm

Figure 3 is bar graph which displays the frequency of the sentiments displayed by the students. By the visualization we can identify that the positive emotions has the higher frequency. Similarly the lowest ones are the disgust and surprise.

Figure 4 represents the accuracies of all the model which were applied on the dataset. The accuracies of depression and sarcasm are depicted side by side with each model. The colour representation makes it easier to visualize the data through graph.

The decision tree model gives out two trees for each depression and sarcasm which are depicted below in Figure 5 and Figure 6

Timeline

Description automatically generated Timeline

Description automatically generated

Figure 5: Decision tree of Depression tree. Figure 6: Decision tree of Sarcasm tree

**CONCLUSION**

As a conclusion, KNN, SVM, Decision trees, Naive Bayes, and Random Forest were among the machine learning algorithms employed in our project on depression analysis and sarcasm analysis among students. The outcomes demonstrated that Random Forest provided the highest accuracy for both the sentiments.

Important ramifications of the project include the ability to spot despair and sarcasm in pupils, which can aid in early diagnosis and assistance. The project's findings can be used to provide focused therapies for pupils who might be depressed or acting sarcastically.

This project still has to be improved in a number of ways, though. To increase the precision of the depressive and sarcastic analysis tasks, it may be necessary to investigate the inclusion of extra contextual variables in the text data, such as sentiment analysis and social network analysis. Also, more study may be done to increase the number and calibre of training datasets, which will help all algorithms perform better.

Overall, the project has demonstrated the potential of machine learning algorithms for student sarcasm and sadness identification, but there is still opportunity for development in both areas. The project's hypothesis is that by increasing the algorithms' accuracy, it will be possible to identify kids more accurately who are depressed or who use sarcasm as a coping method. This will enable better support and treatments for these students.

**REFERENCES**

[1] K. Dashtipour, M. Gogate, A. Adeel, H. Larijani, and A. Hussain, “Sentiment analysis of persian movie reviews using deep learning,” *Entropy*, vol. 23, no. 5, 2021, doi: 10.3390/e23050596.

[2] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, “Depression detection from social network data using machine learning techniques,” *Health Inf Sci Syst*, vol. 6, no. 1, Dec. 2018, doi: 10.1007/s13755-018-0046-0.

[3] “WHO (World Health Organization)”, Accessed: Apr. 06, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[4] “The Daily star”, Accessed: Apr. 05, 2023. [Online]. Available: The Daily Star. (2019). High prevalence of depression and anxiety dis-orders. [online] Available at: https://www.thedailystar.net/editorial/high-prevalence-depression-and-anxiety-disorders-1367674

[5] S. Kumar, M. Gahalawat, P. P. Roy, D. P. Dogra, and B. G. Kim, “Exploring impact of age and gender on sentiment analysis using machine learning,” *Electronics (Switzerland)*, vol. 9, no. 2, Feb. 2020, doi: 10.3390/electronics9020374.

[6] A. Naresh and P. Venkata Krishna, “An efficient approach for sentiment analysis using machine learning algorithm,” *Evol Intell*, vol. 14, no. 2, pp. 725–731, Jun. 2021, doi: 10.1007/s12065-020-00429-1.

[7] M. S. Razali, A. A. Halin, L. Ye, S. Doraisamy, and N. M. Norowi, “Sarcasm Detection Using Deep Learning with Contextual Features,” *IEEE Access*, vol. 9, pp. 68609–68618, 2021, doi: 10.1109/ACCESS.2021.3076789.

[8] I. Ali Kandhro, M. Ameen Chhajro, K. Kumar, H. N. Lashari, and U. Khan, “Student Feedback Sentiment Analysis Model Using Various Machine Learning Schemes A Review,” *Indian J Sci Technol*, vol. 14, no. 12, pp. 1–9, Apr. 2019, doi: 10.17485/ijst/2019/v12i14/143243.

[9] M. Sykora, S. Elayan, and T. W. Jackson, “A qualitative analysis of sarcasm, irony and related #hashtags on Twitter,” *Big Data Soc*, vol. 7, no. 2, 2020, doi: 10.1177/2053951720972735.

[10] H. Yu, E. Lee, and S. B. Lee, “SymBiosis: Anti-Censorship and Anonymous Web-Browsing Ecosystem,” *IEEE Access*, vol. 4, pp. 3547–3556, 2016, doi: 10.1109/ACCESS.2016.2585163.

[11] B. Le and H. Nguyen, “Twitter sentiment analysis using machine learning techniques,” in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2015, pp. 279–289. doi: 10.1007/978-3-319-17996-4\_25.

[12] P. Baid, A. Gupta, and N. Chaplot, “Sentiment Analysis of Movie Reviews using Machine Learning Techniques,” *Int J Comput Appl*, vol. 179, no. 7, pp. 45–49, Dec. 2017, doi: 10.5120/ijca2017916005.

[13] I. A. Kandhro, S. Wasi, K. Kumar, M. Rind, and M. Ameen, “Sentiment Analysis of Student�s Comment by using Long-Short Term Model,” *Indian J Sci Technol*, vol. 12, no. 8, pp. 1–16, Feb. 2019, doi: 10.17485/ijst/2019/v12i8/141741.

[14] A. Mitra, “Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset),” *Journal of Ubiquitous Computing and Communication Technologies*, vol. 2, no. 3, pp. 145–152, Sep. 2020, doi: 10.36548/jucct.2020.3.004.

[15] A. Rahman and M. S. Hossen, “Sentiment Analysis on Movie Review Data Using Machine Learning Approach,” in *2019 International Conference on Bangla Speech and Language Processing, ICBSLP 2019*, Institute of Electrical and Electronics Engineers Inc., Sep. 2019. doi: 10.1109/ICBSLP47725.2019.201470.

[16] J. Singh, G. Singh, and R. Singh, “Optimization of sentiment analysis using machine learning classifiers,” *Human-centric Computing and Information Sciences*, vol. 7, no. 1, Dec. 2017, doi: 10.1186/s13673-017-0116-3.