Criterion-1: Curricular Aspects

Key Indicator – 1.3: Curriculum Enrichment

Metric: 1.3.3

Programme: M.Sc. Informatics

Syllabus	https://www.du.ac.in/uploads/RevisedSyllabi1/Annexure-33.%20M.Sc.%20(Informatics)%20IIC_Syllabus_V1-1a-12_final_20918.pdf
Names of Students undergoing internships	Annexure-I
Sample Research Reports	Annexure-II

Annexure-I

Names of students undergoing internships

Institute of Informatics & Communication
University of Delhi South Campus
Benito Juarez Road, New Delhi - 110021

UDSC

Ph: 011-24157397
E-mail: iicoffice@south.du.ac.in

These are the following interns working under the 06 months internship from Nov 2021 to Nov 2023 in the batches.

Batch I	Batch II	Batch III
Anmol Dogra	Thushari Sathsarani Pahalage Dona	Rohit Kumar Sharma
Rahul Chawla	Vaibhav Mehra	Bharat Aggarwal
Priyal Gupta	Karanpreet Singh	Aashi Ansari
Kunal Sharma	Harshita Sharma	Simran
Shilpi Kumari	Vivek Verma	Shilpi
Abhishek Vishwakarma	Pratibha Dohare	Thahir Ahmaed T
Ausaf Ahmad Ansari	Abhishek Shukla	
Dipali Pawar		

In charge

Annexure-II

Sample Research Reports

Predictive analysis based on feature relevance estimation for survival rate post heart failure using ensemble learners

Kunal Sharma
Institute of Informatics and Communication
University of Delhi South Campus
New Delhi, India
kunal.sharma@iic.ac.in

Rahul Chawla
Institute of Informatics and Communication
University of Delhi South Campus
New Delhi, India
rahul.chawla@iic.ac.in

Unmesh Shukla
Institute of Informatics and Communication
University of Delhi South Campus
New Delhi, India
unmesh.shukla@iic.ac.in

Nitisha Aggarwal
Institute of Informatics and Communication
University of Delhi South Campus
New Delhi, India
nitisha.aggarwal@iic.ac.in

Anil Singh Bafila
Institute of Informatics and Communication
University of Delhi South Campus
New Delhi, India
anil.singh@iic.ac.in

Sanjeev Singh
Institute of Informatics and Communication
University of Delhi South Campus
New Delhi, India
sanjeev@south.du.ac.in

Amit Pundir
Department of Electronics
Maharaja Agrasen College, University of Delhi
New Delhi, India
amitpundir@mac.du.ac.in

Geetika Jain Saxena
Department of Electronics
Maharaja Agrasen College, University of Delhi
New Delhi, India
gsaxena@mac.du.ac.in

***Abstract*— Predicting the chances of survival due to heart failure is a complex and challenging problem. However, nowadays, advanced data science and machine learning techniques can predict such probabilities by learning data patterns with high accuracy and reliability. This study uses supervised machine learning classification algorithms to predict death post heart failure. Multiple strong learners and an ensemble of weak learners are analyzed in this study. Data imbalance is handled using the class weights method. Logistic Regression, Support Vector Machine, Decision Tree, Extra Trees classifier, Random Forest, XGBoost and CatBoost are used with optimization to infer the probability of death due to heart failure. Statistical tests such as Chi-square, ANOVA and phi-k correlation tests, and Random Forest-based feature selection techniques are used. The vote bank method from all feature selection techniques has been used to select the most medically relevant features that classify with the highest accuracy. After evaluating different classification models, the highest accuracy was obtained using Random Forest, XGBoost and CatBoost. This study reports a classification accuracy of 96.67%, which improves 8% over the previously published work.**

Keywords—Heart Failure, Machine Learning, Feature Selection, Ensemble Learning, Hyperparameter Optimization

I. Introduction

Cardiovascular disease (CVD) is the leading cause of heart failure among young and old, accounting for the majority of deaths worldwide, as reported by World Health Organisation (WHO) [1]. Heart failure doesn't mean the heart has failed or stopped functioning. Instead, it means that the heart doesn't pump blood to the body as much as it should. People with diabetes, high blood pressure, high cholesterol, and abnormal pulse rate are more likely to have CVD risk factors. In addition, an unhealthy diet and improper lifestyle can be significant contributing risk factors for heart diseases. A study by WHO (2021) reported that 17.9 million people died from CVD in 2019, amounting 32% of deaths world over and 85% of these were heart attacks and strokes [2]. Heart failure is a severe condition in which the timing of diagnosis is critical. Its early diagnosis and detection of symptoms help doctors decide on a treatment program to delay heart failure or mitigate other associated risk factors.

Machine Learning (ML) has proven to be a suitable and reliable tool to predict the death of patients who have had heart failure or are very likely to have a heart attack. ML methods can identify relevant and significant features, rank them and assign a score to them to predict heart failure [3]. Feature selection techniques analyze important attributes that might directly relate to heart failure [4]. Research is going on worldwide to improve these methods' accuracy, precision, and reliability using fewer highly relevant features.

The present study applied ML algorithms to the Heart Failure Clinical Records dataset as described in section 3. Various conventional supervised ML classification algorithms such as Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM) and MLP, and

978-1-6654-8316-2/22/$31.00 ©2022 IEEE

ensemble techniques such as Random Forest (RF), CatBoost (CB), Extra Trees (ETC) and XGBoost (XGB) were used to predict the survival post heart failure. However, the dataset used has a class imbalance. Negating the majority class bias induced by imbalance during learning is essential. Therefore, the present analysis, uses class weightage to balance its contribution during the training of the models.

II. Literature Survey

Researchers worldwide are paving a path using modern ML techniques to predict the occurrence of heart failure precisely and accurately. For example, Meng et al. [5] studied the low left ventricular ejection fraction (LVEF) for predicting sudden cardiac death in patients with low LVEF. T. Ahmad et al. [6] reported survival analysis of patients who have had heart failure and inferred *age*, *serum_creatinine*, *high_blood_pressure*, *anaemia* and *ejection_fraction* as the significant factors causing mortality. No notable deviations were seen due to *smoking*, *diabetes* and *gender* of the sick individuals. Ali et al. [7], 2019, developed an expert system to predict heart failure by stacking an L_1 regularized SVM and an L_2 regularized SVM. A hybrid grid search algorithm (HGSA) was applied to optimize both models simultaneously and the proposed solution attained an accuracy of 92.22%. Guidi et al. [8] examined a clinical decision support system (CDSS) to predict heart failure. Their research used different classification algorithms, compared performances, and found that RF and CART obtained accuracies of 85.60% and 87.60%, respectively. Parthiban and Srivatsa [9] focused on patients with heart issues and diabetes using features like sex, family heredity, blood pressure, glucose levels, and age for predictive analysis of heart disease. They reported an accuracy of 94.60% using the SVM classifier. Mohan et al. [10] suggested a hybrid model for predicting heart disease using a feature selection method and obtained an accuracy of 88.70%. Mohammed W. Akhter et al. [11] found that Renal Insufficiency (RI), due to the elevated *serum creatinine* (> 1.5 mg/dl), is commonly found in patients that have been hospitalized with decompensated heart failure. A. Ishaq et al. [12] analyzed the dataset imbalance problem using SMOTE and found relevant features using the random forest feature selection method. Various classification algorithms were reported with maximum accuracy of 92.62% using the Extra Trees classifier. D. Chicco et al. [3] reported feature relevance and found *ejection fraction* and *serum creatinine* to be the most significant attributes in predicting death due to heart failure.

Extensive research has been reported in the literature highlighting the approaches that performed well in predicting outcomes related to heart diseases or cardiac failure Table I.

III. Proposed Methodology

The current study compares different machine learning algorithms for predicting death due to heart failure. The following steps were carried out for this study and the detailed workflow is shown in Fig. 1.

- At the onset, the EDA of the dataset was performed with DEATH_EVENT as the target attribute. Insightful visualizations and other significant aspects of the data were obtained.
- Post EDA, data preprocessing was done wherein the original dataset was split into train and test subsets. After splitting, feature scaling was performed to center the data around zero and scale it to unit variance.
- Chi-square, ANOVA and ϕ_k correlation tests and the Random Forest-Based Feature Selection technique were used to determine the relevance of each feature for predicting the target attribute. According to the calculated relevance values, the best feature sets indicated by each algorithm were identified. The intersection of these feature sets was obtained as the reduced feature set and classification performance metrics were determined for this set.
- Class weights were assigned to the two target classes, namely, DEATH_EVENT = 0 and DEATH_EVENT = 1, to compensate for the bias of training models towards the majority class.
- LR, SVM (kernel Linear, Poly, RBF), DT, ETC, RF, XGB and CB were evaluated for classification. Grid-search cross-validation was used to optimize hyperparameters of the ML models. Furthermore, maximum accuracy, precision, recall, f1-score and AUC-ROC were recorded.

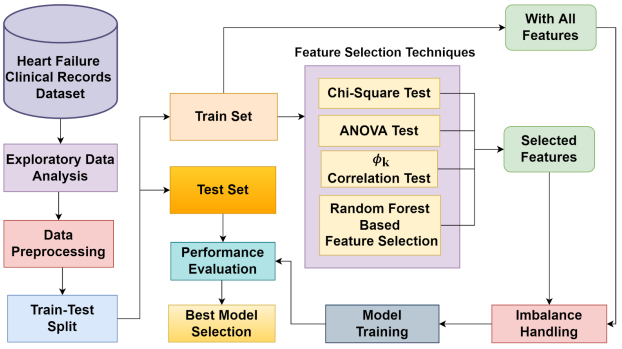


Fig. 1. Workflow of the proposed study.

A. Dataset Description and Exploratory Data Analysis

The Heart Failure Clinical Records dataset, comprising the medical records of 299 patients (194 - male and 105 - female) with 13 clinical features [13], was used in this study. The target attribute, the DEATH_EVENT, consists of classes 0 and 1, corresponding to the patients surviving and not surviving the heart failure, respectively. In addition, the target comprises imbalanced class labels, i.e., the number of instances of class 0 is 203, whereas that of class 1 is 96. The detailed description of dataset features is given in Table II.

The EDA showed that the dataset consists of six categorical and seven numerical attributes. However, it consists of no duplicate entries and no missing values. Still, most of the features are skewed and have outliers. The detailed descriptive statistical analysis of the numerical features of the dataset is shown in Table III. From Fig. 2, it can be seen that DEATH_EVENT is correlated with various other attributes. It is also observed that *time* and *ejection_fraction* show a negative correlation, whereas *serum_creatinine* shows a positive correlation with DEATH_EVENT. Furthermore, various correlation measures such as Pearson's, ϕ_k, Cramer's v and Kendall's coefficients were also observed in the form of heatmaps.

B. Data Preprocessing

The dataset was divided into a 70:30 train-test ratio and standard scaling was used to overcome the effect of data points with varying magnitudes. Standardization is a scaling technique where scaling for all the values of independent variables are done which is to be centered around zero with a standard deviation equal to 1. It is done to normalize the range of attributes or independent variables before training. Standard scaling is governed by (1).

$$S_i = (x_i - \mu) / \sigma \quad (1)$$

Here, S_i is the scaled value and x_i is the value of an attribute x at i^{th} instance, μ is the mean and σ is the standard deviation of the attribute.

TABLE I. Related Work On Heart Failure And Cardiac Disease

Authors	Problem	Method	Dataset	Accuracy (%)
Chicco et al. [3]	Prediction of survival and to rank features based on risk factors	RF	[13]	74.00
Ali et al. [7]	SVM based expert system to predict heart failure	L1 and L2 Regularized SVM and HGSA	[13]	92.22
Guidi et al. [8]	Clinical decision assistance system for heart failure prediction and analysis	RF and CART	[13]	87.60
Parthiban et al. [9]	People with diabetes with heart issues who are likely to have heart diseases	SVM	Diabetics' Diagnosis Data	94.60
Mohan et al. [10]	Prediction of CVD using a different combination of features	Hybrid RF with Linear model	[13]	88.00
Ishaq et al. [12]	To find useful features and improve accuracy for CVD patients' survival	SMOTE and Random Forest Feature selection with ETC	[13]	92.62
Oladimeji et al. [14]	Prediction of survival	WEKA open source software	[13]	83.17
Kucukakcali et al. [15]	Classification and data mining methods	Relational Classification methods	[13]	86.66
Parisi et al. [16]	Reliability function analysis for SVM and MLP	SVM and MLP with m-arcsinh	[13]	88.00
Gürfidan et al. [17]	Mortality associated with heart disease	SVM	[13]	83.00
Sanni et al. [18]	To find important features and to boost accuracy for CVD patients survival	DT	[13]	85.33
Rahayu et al. [19]	Predict the survival due to heart failure	SMOTE and Resampling technique in RF	[13]	94.31
Sahu et al. [20]	Feature relevance study for Cardiovascular risk assessment using data mining	Using K-fold cross validation and PCA (LR)	Heart Disease Data	85.11

C. Feature Selection Methods

Feature selection is one of the crucial steps in ML, which significantly affects the performance of a model. It can reduce overfitting, improve accuracy and reduce training time for the model. Feature relevance estimation was done to find an effective subset of attributes from the whole dataset to predict death due to heart failure using different techniques.

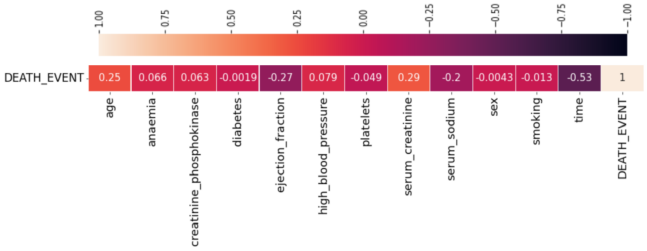


Fig. 2. Correlation coefficients of highly correlated features with the target.

The estimation techniques ranked features with relevance scores and reduced the dataset's dimensionality while keeping the set's most relevant attributes. Chi-Square test, ANOVA test, ϕ_k Correlation Test and Random Forest-Based Feature Selection (RFBFS) were used for relevance estimation and ranking.

TABLE II. Dataset Description

Name	Description
age	Patient's Age
anaemia	Deficiency of RBC or hemoglobin
high_blood_pressure	Whether the patient has hypertension
creatinine_phosphokinase	Creatinine phosphokinase enzyme level in the blood
diabetes	Whether the patient has diabetes
ejection_fraction	Blood leaving the heart at each contraction (in percentage)
sex	Gender of the patient
platelets	Platelets in the blood
serum_creatinine	Creatinine Level in the blood
serum_sodium	Sodium Level in the blood
smoking	Whether the patient smokes
time	Follow-up period
(target) *DEATH_EVENT*	Patients' death during the follow-up

TABLE III. Descriptive Statistics (S1-age, S2 -*creatine_phosphokinase,* S3-*ejection_fraction,* S4-*platelets,* S5-*serum_creatinine,* S6-*serum_sodium,* S7-*time)* (D – Standard Deviation, M – Mean, K – Kurtosis, S - Skewness)

	S1	S2	S3	S4	S5	S6	S7
D	11.9	970.3	11.8	97804.2	1.0	4.4	77.6
M	60.8	581.8	38.0	263358.0	1.4	136.6	130.3
K	-0.2	25.1	0.0	6.2	25.8	4.1	-1.2
S	0.4	4.5	0.6	1.5	4.5	-1.0	0.1

D. Handling Class Imbalance with Class Weights Method

Class imbalance occurs quite frequently in classification tasks. The higher number of occurrences of the instances of one class compared to another may result in a bias or skewness towards the majority class in the model. However, this bias can be avoided by assigning fitting weights to both the majority and the minority classes. Furthermore, specifying the class weights explicitly for an unbalanced dataset manipulates the penalties in the model's cost function, diminishing the biased effect of class imbalance on the model during training.

E. Classification Algorithms

LR, SV, DT, RF, ETC, XGB and CB were used for classifying the dataset instances into dead (1) and survived (0) classes. Hyperparameter optimization for each model can be done using different approaches such as manual hyperparameter tuning, grid search, random search and Bayesian optimization. In this study, manual search and grid search optimization techniques were used. The grid search technique passes a grid of hyperparameters to the model under different permutations and combinations during training and determines performance parameters using cross-validation.

IV. Results and Discussion

The results reported and discussed in this section include feature relevance estimations, performance improvement by imbalance handling and model performance estimation.

A. Feature Relevance

In this study, three tests were implemented. The Chi-square and ANOVA tests were used to find the relevant features on two sets of independent attributes. Chi-square was used to identify relevant categorical variables, while ANOVA identified relevant numerical variables. Further, RFBFS and ϕ_k correlation tests were used to estimate the relevance of all features.

a) Chi-Square and ANOVA tests: The following feature subsets (Subsets 1 and 2) represent categorical and numerical features, respectively.

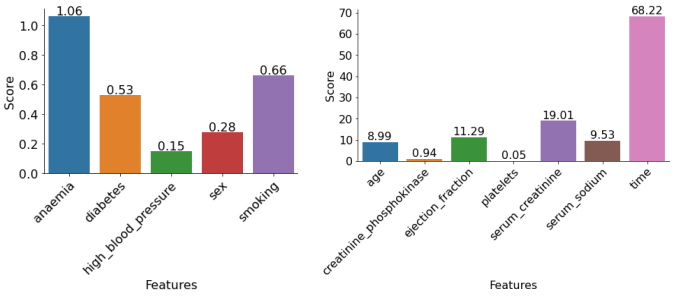
Subset 1: categorical attributes
anaemia, diabetes, high_blood_pressure, sex, smoking

Subset 2: numerical attributes
age, creatinine_phosphokinase, ejection_fraction, platelets, serum_creatinine, serum_sodium, time

The Chi-square test was performed for Subset 1 and the ANOVA test was carried out for Subset 2. The relevance scores derived from the respective tests for the corresponding subsets are shown in Fig. 3. It is evident from the scores of both the tests that *anaemia, time, serum_creatinine*, *ejection_fraction*, *age* and *serum_sodium* had relatively higher scores than the rest of the features.

b) Random Forest-Based Feature Selection (RFBFS): RF was used to select features using the scikit-learn SelectFromModel function. The threshold considered in this technique was evaluated based on the mean of the scores assigned by the RF classifier to each feature after training, whose value was 0.08. Fig. 4 shows the scores for all the features that are obtained from the RFBFS technique. All those features whose scores are greater than 0.08 were selected. The selected attributes are *age, creatinine_phosphokinase, ejection_fraction, platelets, serum_creatinine* and *time.*

c) ϕ_k Correlation Test: This test was implemented using the *phik* library in Python, which was also used to find relevant features and plot heatmaps, as shown in Fig. 5. *The correlation values, in Fig. 5, show that* time, serum_sodium, serum_creatinine, ejection_fraction *and* age *have a good correlation score with the target attribute DEATH_EVENT.*



(a) (b)

Fig. 3. Feature scores for (a) Chi-Square (b) ANOVA tests.

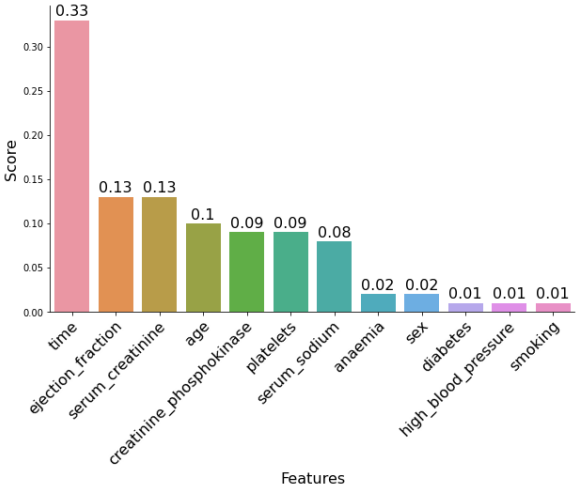


Fig. 4. Scores calculated from RFBFS.

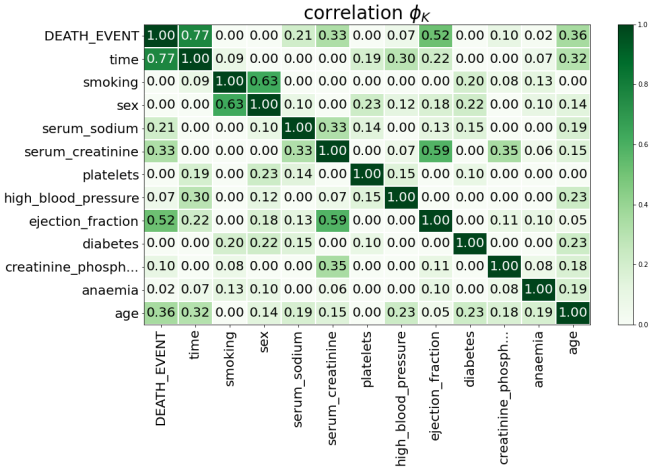


Fig. 5. Heatmap for ϕ_k correlation.

Hence, the features found to be relevant after verification using correlation and statistical significance tests were *time, serum_sodium, serum_creatinine, ejection_fraction* and *age*.

d) *The intersection of feature selection techniques:* Application of the various feature selection techniques yielded relevant features for predicting death due to heart failure. The common significant attributes were *age, ejection fraction, serum_creatinine,* and *time*. These features were concluded to be the most relevant for predicting death due to heart failure. The feature selection process adopted for the present study is shown in Fig. 7.

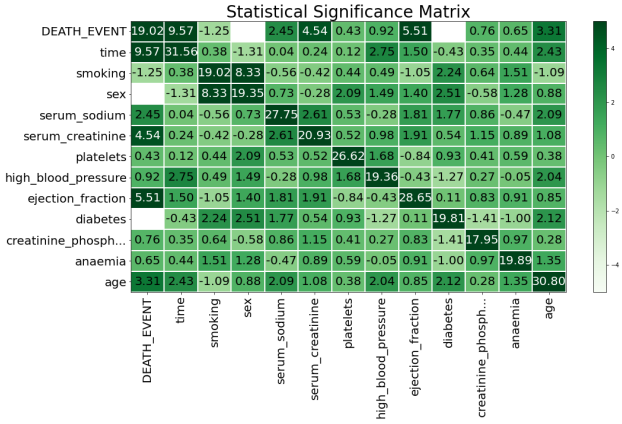


Fig. 6. Significance Matrix using ϕ_k correlation.

B. *Imbalance Handling and Prediction*

The dataset considered in the present study has an imbalance towards the class of the patients that survived after heart failure, introducing a bias in learning. However, predicting both classes, majority and minority, with high precision and accuracy is equally important; hence the *class weights* were considered for training all the models. This method ensures the assignment of appropriate fitting weights to each class during training, which further removes the bias for the majority class while training.

The percentage of 'class 0 (Survived)' was 66.50% in the train set and 71.11% in the test set, while 'class 1 (Died)' was 33.49% in the train set and 28.88% in the test set. LR, SVM (with the kernel as Linear, Poly, RBF), DT, ETC, RF, CB and XGB were trained. Different evaluation metrics were obtained considering these percentage weights.

C. *Model Training and Evaluation*

In this study, fine-tuning of the internal hyperparameters of all ML models was done. For bagging and boosting algorithms, the parameters *max_depth* and *random_state* were tuned. GridSearchCV was used for tuning the rest of the classifiers. All models were evaluated post-optimization of hyperparameters, using accuracy, Matthew's Correlation Coefficient (MCC), AUC-ROC, precision, recall and f1-score as the performance metrics. RF, XGB and CB were found to outperform other algorithms and have the maximum accuracy. RF, trained only on features selected, i.e., *age, ejection_fraction, serum_creatinine* and *time*, gave the best classification accuracy and other performance metrics while XGB and CB performed best with all features. Results for the same are shown in Table IV and learning curves are shown in Fig. 8 and Fig. 9. The best parameters for RF, XGB, CB are listed in Table V.

The maximum accuracy obtained in the study was 96.67%, using class weights. In addition, different profiles of evaluation metrics such as Area Under Curve (AUC)-Receiver Operating Characteristics (ROC), Cumulative Accuracy Profile (CAP) and confusion matrices were obtained to determine the ML models' efficiency to classify

the target. Fig. 10 shows confusion matrices for RF, XGB, CB, ETC created to calculate the AUC-ROC, Precision, Recall and F1-Score (Table IV).

ROC is an evaluation metric for binary classification in target attributes. Its curve is used as a probability curve that define the True Positive Rate (TPR) versus False Positive Rate (FPR), where TPR is TP/(TP+FN) and FPR is FP/(FP+TN).

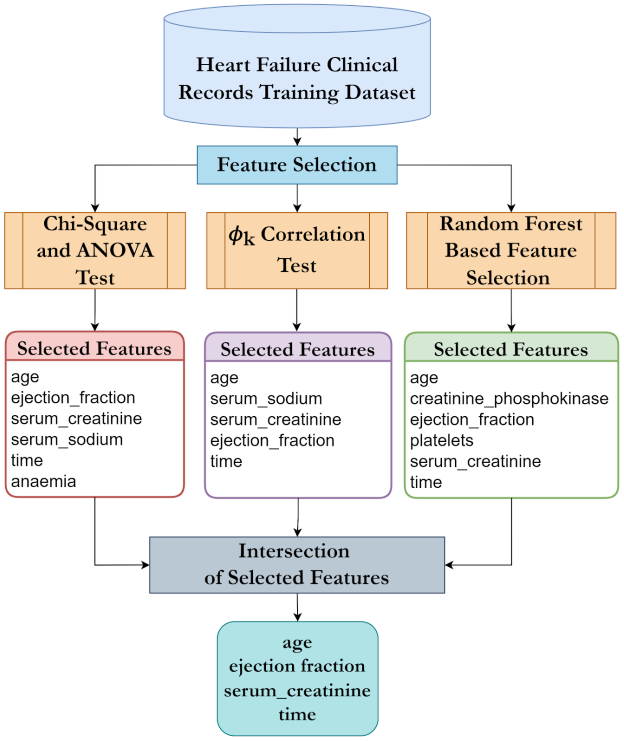
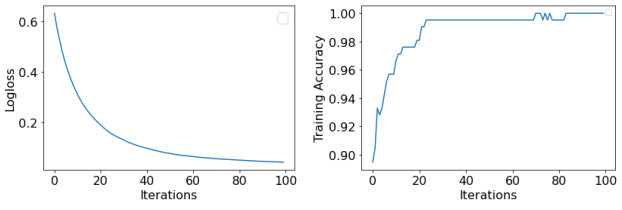
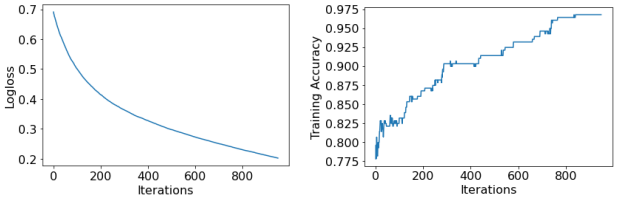


Fig. 7. Workflow for Intersection of Feature Selection Techniques.



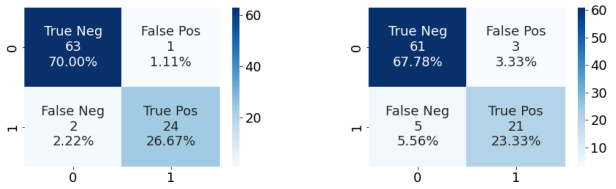
(a) (b)

Fig. 8. (a) XGBoost (XGB) log loss versus iteration (b) XGBoost (XGB) training accuracy versus iteration.



(a) (b)

Fig. 9. (a) CatBoost (CB) log loss versus iteration (b) CatBoost (CB) training accuracy versus iteration.



(a) (b)

Fig.10. Confusion Matrix (a)RF/XGB/CB (b) ETC.

The CAP curve is a simple and robust method widely used to visualize the discriminative power of the model and compare machine learning classifiers. The cumulative number of positive results is represented on the y-axis, while the x-axis represents the corresponding cumulative number of classifying parameters.

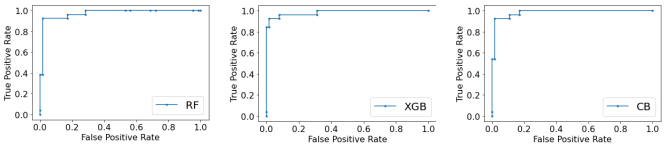
Any model can be analyzed by comparing its CAP curve with perfect and random CAP curves. The maximum number of positive outcomes achieved directly result in the perfect CAP curve. In contrast, the random CAP curve has the positive outcomes distributed equally. A model is good if its CAP curve lies between perfect and random CAP curves. The model is better if it is closer to the perfect curve than the random curve.

TABLE IV. MODEL EVALUATION METRICS (A – ACCURACY, MCC – MATTHEWS CORRELATION COEFFICIENT, AUC-ROC – AREA UNDER THE CURVE – RECEIVER OPERATING CHARACTERISTIC, P – PRECISION, R – RECALL, F– F1-SCORE)

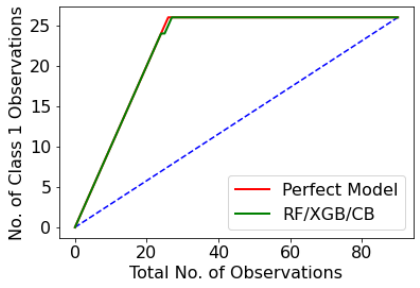
Model	A(%)	MCC	AUC-ROC	P	R	F
RF	96.67	0.92	0.95	0.97	0.97	0.97
LR	86.67	0.68	0.85	0.87	0.87	0.87
DT	90	0.75	0.85	0.9	0.9	0.9
SVM (Linear)	83.33	0.61	0.81	0.83	0.83	0.83
SVM (Poly)	81.11	0.52	0.74	0.81	0.81	0.81
SVM (RBF)	83.33	0.6	0.8	0.83	0.83	0.83
ETC	91.11	0.78	0.88	0.91	0.91	0.91
XGB	96.67	0.92	0.95	0.97	0.97	0.97
CB	96.67	0.92	0.95	0.97	0.97	0.97

TABLE V. BEST MODEL HYPERPARAMETERS

Model	Hyperparameters
RF	max_depth = 3, random state = 32, class_weight = {'0':2.11, '1':1}
XGB	max_depth = 9, scale_pos_weight = 2.11
CB	max_depth = 4, class_weight = {'0':2.11, '1':1}



(a)



(b)

Fig. 11. (a) ROC curves showing the tradeoff between the true positive rate and false-positive rate for RF/XGB/CB (b) CAP curves for RF/XGB/CB.

The ROC and CAP profiles computed for the best-trained models RF, XGB, CB on the Heart Failure Clinical Records Dataset [13] are shown in Fig. 11.

V. Conclusion

The maximum accuracy of predicting death post heart failure was 96.67% for RF, XGB and CB algorithms. Different hyperparameter optimizations were implemented to achieve the best result. Furthermore, the dataset imbalance was handled using class weights by giving equal importance to the majority and minority classes during the training phase. Comparison of the results computed in present work with the previous studies (Table VI) on the same dataset based on classification accuracy is summarized in the results and discussion section, which shows that the results in this study are better than most of the previous work. Since the proposed model predicts the classes of unseen data with very high accuracy, it can be of immense use in biomedical applications. However, the small sample size, i.e., 299 patient records, and the limited geographic localization of the patients are the factors that will restrict the generalizability of this study. Therefore, the proposed model cannot be applied globally without tuning it for more varied data from worldwide sources. Hence, the limitation of the current study can act as a scope of future research in this field to develop a more robust prediction model. However, for future work if more data or patients records of different geographic location with same attributes can be achieved than the model can lead to this limitation by building a more robust model for predicting death post heart failure.

TABLE VI. Comparative Analysis of Results with Previous Studies

Method	Accuracy (%)	Reference
Classification algorithms to predict survival of patients post heart failure, using WEKA open source software (after feature selection)	83.17	[14]
Predicting mortality due to heart failure using association rules (relational classification methods)	86.66	[15]
Predicting mortality using m-arcsinh algorithm	88.00	[16]
Analyzing performance metrics using ML algorithms	85.33	[18]
Predicting mortality after applying resampling technique on heart failure dataset using ML algorithms	94.31	[19]
Present study	**96.67**	-

Acknowledgment

The authors thank the Institute of Eminence (IoE), the University of Delhi, for the FRP scheme research grant and the Ministry of Education (MoE), Government of India initiative Project SAMARTH, at University of Delhi South Campus (UDSC), for their support.

References

[1] Cardiovascular Disease, in Health-section, World Health Organization, 2021. Accessed on: Oct. 25, 2021. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases.

[2] Cardiovascular disease, in Newsroom/Fact Sheet/detail, World Health Organization (WHO), 2021. Accessed on: Oct. 25, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[3] D. Chicco and G. Jurman, "Machine-learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," in BMC Med Inform Decis, vol. 20, 2020, pp. 1-16.

[4] N. Hasan and Y. Bao, "Comparing different feature selection algorithms for cardiovascular disease prediction," in Health Technol, vol. 11, 2021, pp. 49–62.

[5] F. Meng, Z. Zhang, X. Hou, Z. Qian, Y. Wang, Y. Chen, Y. Wang, Y. Zhou, Z. Chen, X. Zhang and J. Yang, "Machine learning for prediction of sudden cardiac death in heart failure patients with low left ventricular ejection fraction: study protocol for a retrospective multicentre registry in China," in BMJ, vol. 9, 2019, pp. e023724.

[6] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab and M. A. Raza, "Survival analysis of heart failure patients: A case study," in PLOS One, vol. 12, 2017, pp. e0181001.

[7] L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R. Nour and S. A. C. Bukhari, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure," in IEEE Access, vol. 7, 2019, pp. 54007-54014.

[8] G. Guidi, M. C. Pettenati, P. Melillo and E. Iadanza, "A machine learning system to improve heart failure patient assistance," in IEEE J Biomed Health, vol. 18, no. 6, 2014, pp. 1750-1756.

[9] G. Parthiban and S. K. Srivatsa, "Applying machine learning methods in diagnosing heart disease for diabetic patients," in International Journal of Applied Information Systems (IJAIS), vol. 3, no. 7, 2012, pp. 25-30.

[10] S. Mohan, C. Thirumalai and G. Srivastav, "Effective heart disease prediction using hybrid machine learning techniques," in IEEE Access, vol. 7, 2019, pp. 81542-81554.

[11] M. W. Akhter, D. Aronson, F. Bitar, S. Khan, H. Singh, R. P. Singh, A. J. Burger and U. Elkayam, "Effect of elevated admission serum creatinine and its worsening on outcome in hospitalized patients with decompensated heart failure," in AM J Cardiol, vol. 94, no. 7, 2004, pp. 957-960.

[12] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara and M. Nappi, "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," in IEEE Access, vol. 9, 2021, pp. 39707-39716.

[13] D. Dua and C. Graff, "Heart Failure Clinical Records Dataset," in UCI Machine Learning Repository, Centre for Machine Learning and Intelligent systems, 2019, last accessed: Oct. 10, 2021. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

[14] O. O. Oladimeji and O. Oladimeji, "Predicting survival of heart failure patients using classification algorithms," in Journal of Information Technology and Computer Engineering (JITCE), vol. 4, 2020, pp. 90-94.

[15] Z. Kucukakcali, I. B. Cicek, E. Guldogan and C. Colak, "Assessment of associative classification approach for predicting mortality by heart failure," in The Journal of Cognitive Systems, vol. 5, no. 2, 2020, pp. 41-45.

[16] L. Parisi, "m-arcsinh: An efficient and reliable function for SVM and MLP in Scikit-learn," CoRR, vol. arXiv:2009.07530, 2020.

[17] R. Gürfidan and M. Ersoy, "Classification of death related to heart failure by machine learning algorithms," in Adv Art Int, vol. 1, no. 1, 2021, pp. 13-18.

[18] R. R. Sanni and H. S. Guruprasad, "Analysis of performance metrics of heart failure patients using Python and machine learning algorithms," in Global Transitions Proceedings, vol. 2, 2021, pp. 233-237.

[19] S. Rahayu, J. Purnama, A. Pohan, F. Nugraha, S. Nurdiani and S. Hadianti, "Prediction of survival of heart failure patients using random forest," in Pilar Nusa Mandiri: Journal of Computing and Information System, vol. 16, 2020, pp. 255-260.

[20] A. Sahu, G. M. Harshvardhan, M. K. Gourisaria, S. S. Rautaray and M. Pandey, "Cardiovascular risk assessment using data mining inferencing and feature engineering techniques," in Int J Inf. Tecnol, vol. 13, 2021, pp. 1–13.

Transfer Learning based Cardiac Murmur Detection in Phonocardiogram Signals using Spectrograms

Pratibha Dohare[a], Unmesh Shukla[b], Diptadeep Bhattacharjee[c], Sanjeev Singh[b], Amit Pundir[d] and Geetika Jain Saxena[d]*

[a]*Cluster Innovation Centre, University of Delhi, Delhi, India;* [b]*Institute of Informatics and Communication, University of Delhi, Delhi, India;* [c]*National Institute of Technology Silchar, Assam, India;* [d]*Department of Electronics, Maharaja Agrasen College, University of Delhi, Delhi, India*

*** Corresponding author details**
Name - Geetika Jain Saxena
Email - gsaxena@mac.du.ac.in
Affiliation - Department of Electronics, Maharaja Agrasen College, University of Delhi, Delhi, India

Transfer Learning based Cardiac Murmur Detection in Phonocardiogram Signals using Spectrograms

This study focuses on the detection of cardiac murmurs, which are indicative of serious heart conditions such as cardiovascular diseases (CVD). Manual interpretation of phonocardiogram (PCG) signals is often difficult due to the noise from the external background and the time required by a human expert, which results in the need for an automated system to detect murmurs. This study used transfer learning architectures to detect the presence of murmurs in PCG signals. The murmur detection method proposed in this study comprises preprocessing techniques to remove signal noise, and feature extraction techniques that generate spectrograms to convert input signals to suitable inputs for the transfer learning architectures. The techniques of Short-Time Fourier Transform (STFT), Mel-Frequency Cepstral Coefficients (MFCC), and Continuous Wavelet Transform (CWT) were applied on Physionet's CirCor Digiscope PCG dataset to generate spectrograms and compared for feature extraction. VGG16, VGG19, ResNet50, and InceptionV3 models, were trained on these spectrograms for binary classification. Fourth-order Butterworth bandpass filter, with a cutoff frequency range of 20-400 Hz, used with Savitzky-Golay filtering gave the best results. As compared to other combinations, the CWT Spectrogram and VGG19 combination performed best for murmur detection with an accuracy of 89.44%. Different combinations of spectrograms and transfer learning architectures performed better on performance metrics of precision, recall, F1-score, and ROC-AUC. The study found that transfer learning models, combined with bandpass filtering, provide reasonable accuracy for detecting murmurs using the CWT spectrograms of PCG signals.

Keywords: Heart sound classification, transfer learning, spectrogram, phonocardiogram, machine learning

1. Introduction

Cardiovascular Diseases (CVDs) are the cause of about 32% of overall deaths worldwide [1]. The damage due to CVDs can be reduced through combinations of lifestyle changes and medical interventions aimed at reducing the risk of further disease progression and reducing the risk of adverse events such as myocardial infarction [2, 3]. Hence, it is critical to detect the risk or presence of CVDs at an early stage.

Cardiac auscultation is a crucial non-invasive diagnostic technique of listening to the sound patterns produced by the heart, usually using a stethoscope or phonocardiography equipment. It is used by medical experts to assess functions of the heart. Murmurs are abnormal sounds heard during a cardiac auscultation made by heart during opening and closing of heart valves. They are produced by turbulent blood flow caused by temporary or permanent conditions. Due to the capability of murmurs to suggest the presence of CVDs, they are one of the primary sound patterns that are required to be detected by doctors and physicians during the process of auscultation. Though murmurs do not necessarily imply presence of a CVD, they could be indicators of the early stage of CVDs.

Phonocardiogram (PCG) signals are used to carry these sound patterns. Sub-audible sounds and murmurs are recorded in a PCG signal. As shown in figure 1, the PCG signal can be divided into different parts, each corresponding to specific events and intervals during the heart's functioning. S1 is the first heart sound in the cardiac cycle and is often referred to as the "lub" sound. It is produced by the closure of the atrioventricular valves (mitral and tricuspid valves) during ventricular systole (contraction). It marks the beginning of the systolic phase of the cardiac cycle, indicating the initiation of ventricular contraction and the ejection of blood from the heart's ventricles into the arteries. S2 is the second heart sound in the cardiac cycle and is

commonly referred to as the "dub" sound. It is caused by the closure of the semilunar valves (aortic and pulmonary valves) during ventricular diastole (relaxation). It marks the end of the systolic phase and the beginning of the diastolic phase, signifying the cessation of ventricular contraction and the onset of ventricular relaxation and filling. The systolic interval represents the duration between the onset of S1 and the end of S2. It corresponds to the period when the ventricles are contracting (systole), pumping blood out of the heart into the pulmonary artery (right ventricle) and aorta (left ventricle). During the systolic interval, arterial pressure increases as blood is expelled from the ventricles. The diastolic interval refers to the time between the end of S2 and the onset of the next S1. It represents the period when the ventricles are relaxing (diastole), allowing them to fill with blood from the atria. During the diastolic interval, the heart is in a resting phase, and blood returns to the heart to prepare for the next cardiac cycle.

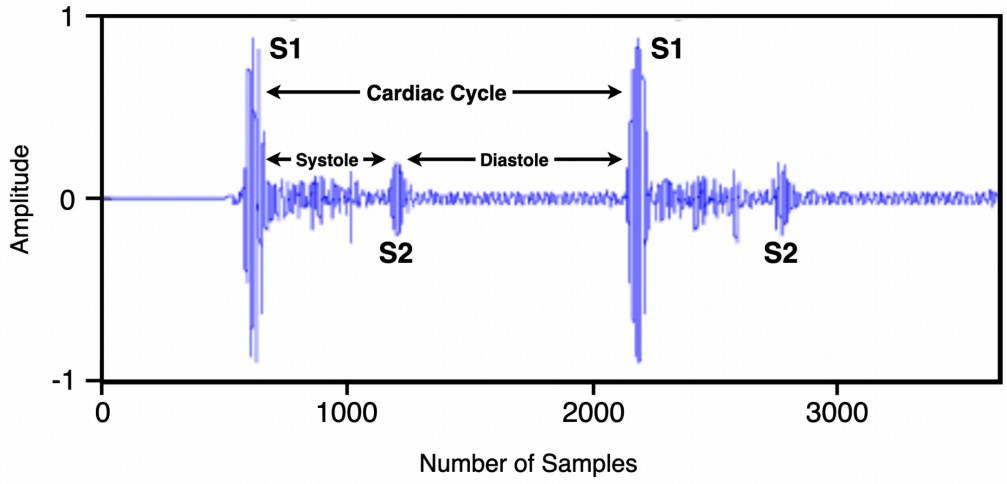


Figure 1. Different parts of the PCG signal

Signal acquisition

A PCG signal comprises a high quality recording of heart sounds, typically recorded by a phonocardiograph that uses a stethoscope. These recordings can either be done using specialized mechanical or electronic stethoscopes, or through body worn sensors. The current study uses

PCG signals recorded using an electronic stethoscope. This type of stethoscope also has the ability to amplify and filter the sounds, thereby enabling easy detection and analysis of murmurs. The signals were obtained from one or more auscultation locations - Pulmonary Valve (PV), Aortic Valve (AV), Mitral Valve (MV), Tricuspid Valve (TV), and Others (Phc). The location and duration of the recordings vary among patients.

Suitability of PCG signals for diagnosis

In terms of ease of data collection, since phonocardiography is a non-invasive diagnostic technique, it is relatively easier to collect data using it when compared with invasive techniques. Additionally, phonocardiography setup is simpler than other non-invasive techniques such as electrocardiography (ECG) that requires placement of multiple electrodes on the human body. Therefore, due to the ease of recording PCG signals and the common use of these signals in primary diagnosis of the heart, building an automated machine learning (ML) based decision support system that uses murmurs for diagnosis of heart patients makes sense.

Presence and absence of murmurs in PCG signals

PCG inherently provides 1D signals in the time domain. A medical expert usually analyzes these signals in the time domain and infers the presence or absence of murmurs by looking for specific waveforms (Figure 2). However, the process of manual inspection to detect murmurs is time consuming and prone to human errors.

Use of Transfer Learning architectures for murmur detection

A human medical expert directly analyzes 1D PCG signals for diagnosis. However, it takes observation of longer periods of such signals to infer consistent presence of specific waveforms

denoting murmurs. Hence, designing an ML-based automated system for murmur detection in 1D signals requires the extraction of information of highest importance from these signals that can be used to make faster and more reliable decisions than human experts. Since waveforms are characterized by time and frequency, it is often better to transform the signal data to 2D representations that highlight both time and frequency information. Two-dimensional spectrograms of 1D signals are examples of such 2D representations that provide localization in the time and frequency domains. These spectrograms can be obtained by applying techniques such as Fast Fourier Transform (FFT), Short-Time Fourier Transform (STFT), Continuous Wavelet Transform (CWT), and Mel-Frequency Cepstral Coefficients (MFCC) on preprocessed PCG signals. Since transfer learning architectures have already been designed to extract meaningful features from complex images, it is wise to use them for classification of 2D spectrograms. In accordance with the color map used to create the spectrograms, the range of colors present in them signify the different magnitudes of signal frequencies - brighter (like yellow, pink) and darker (like violet) colors denote higher and lower frequencies, respectively. Figure 3, figure 4, and figure 5 show the visual differences in the CWT, MFCC, and STFT spectrograms of PCG signals with and without murmur, respectively.

Information content in spectrograms

In the CWT spectrogram (here, scalogram), time and frequency are represented by the two axes whereas colors denote the magnitude of the frequency components. It is evident from figure 3 that more frequencies are present in PCG signals with murmurs as compared to PCGs without murmur. In the first and last second of signals, high magnitude frequency components are present in more amounts for PCG signals with murmur. The MFCC spectrogram uses a nonlinear

frequency scale based on the Mel scale which is more aligned with the human perception of sound. The x-axis represents time and the y-axis represents the MFCC coefficients. Figure 4 shows that more frequencies of higher magnitudes of MFCC coefficients are present in PCG signals with murmurs. STFT spectrogram is another time-frequency representation of a 1D signal. It is obtained by dividing the signal into small overlapping windows and performing a Fourier transform on each window. The resulting spectrogram shows how the frequency content of the signal changes over time. The x and y axes denote time and frequency, respectively whereas the color intensity denotes the magnitude of frequency content. PCG signals with murmurs have more frequencies as compared to those without murmurs as shown in figure 5.

Rationale for choice of Spectrograms

STFT provides a good balance between time and frequency resolution that makes it suitable for analyzing stationary components in heart sounds. It is a well-established method for heart sound analysis [4-6]. MFCCs have also proven to be effective when applied to heart sound analysis and other related audio tasks [7-9]. The use of frequency bands in the Mel-scale allows them to closely approximate the human auditory system's response, thereby indirectly modeling the ear's sensitivity to changing frequencies. Therefore, employing MFCCs as features is especially well-suited for simulating cardiac auscultation activity. CWT offers excellent time-frequency localization, making it well-suited for capturing transient events and non-stationary characteristics in heart sounds, such as murmurs as evidenced by recent studies [10-12]. Its ability to perform multiresolution analysis proves advantageous when heart sounds contain components at various frequency scales. Furthermore, CWT can be particularly valuable in identifying specific frequency patterns related to abnormalities in heart sounds. Other time-frequency transformation methods either have lesser support from results published in

previous studies as compared to STFT, MFCC, and CWT, or they have inherent drawbacks that make them intuitively relatively weaker for heart sound classification. For instance, CQT (Constant-Q Transform) offers better frequency resolution for low frequencies. However, its exponential frequency spacing might not be as well-suited for heart sound analysis, where different frequency components are relevant as it is for musical applications. Similarly, relative to the three methods used in this study, there is very little evidence in existing literature for the use of gammatone filter banks on heart sounds.

1.1 Motivation

An automated system for the early detection of murmurs can lead to timely diagnosis and treatment of heart diseases or other cardiac conditions, which can improve patient outcomes and reduce healthcare costs. The levels of expertise of human analyzers or listeners of heart sounds may be different, which can lead to inconsistent subjective interpretations about the presence and types of murmurs. On the contrary, a deep learning (DL) based solution for murmur detection is more likely to be objective and consistent. PCG signals recorded for the diagnoses of large durations of time lead to large amounts of data. As compared to manual analysis, an automated solution saves time and resources, thereby being quicker and more accurate. An ML based solution can always be updated through retraining when new signal recordings are available.

1.2 Related works

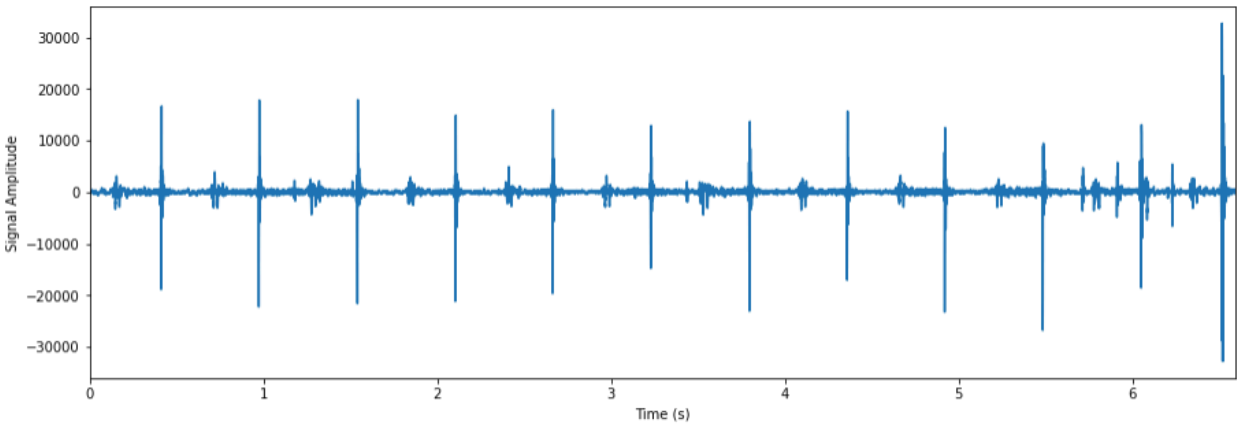
Traditional ML models, transfer learning, and DL techniques have been extensively utilized to extract meaningful information from cardiac sounds in recent times [13-18]. Furthermore, fundamentally different preprocessing techniques and classification algorithms have also been

applied for heart sound classification [16-17]. Spectrograms generated using MFCC, CWT, and STFT techniques have been used to detect pathological heart murmurs [13], [14], [17-18].

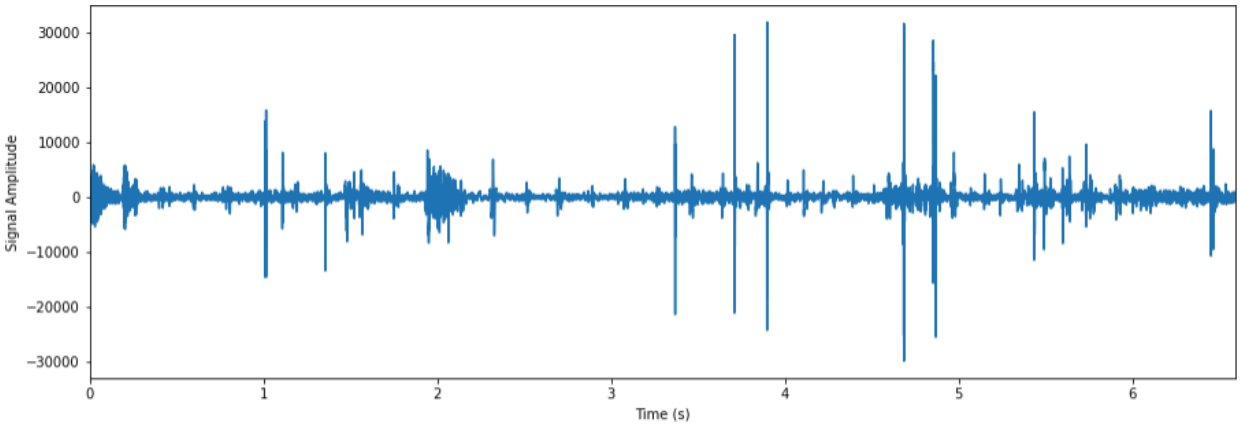
The study in [13] explored the feasibility of using transfer learning to detect heart murmurs from PCG recordings. Transfer learning models like VGG16, VGG19, and ResNet50 were used in this study to classify spectrograms of heart sounds from the PASCAL CHSC database with the highest accuracy of 87.65%. A convolutional neural network (CNN) was used in [14] to classify STFT spectrograms of the PCG dataset from PhysioNet/CinC 2016 [19] and PASCAL 2011 [20] as normal or abnormal with 97.52% accuracy. The study in [15] used a CNN-Bidirectional Long Short-Term Memory (CNN-BiLSTM) network to detect valvular heart diseases from 1D PCG signals with 87.31% accuracy on the PhysioNet/CinC 2016 dataset [19]. DL models were utilized by the study in [16] to obtain a sensitivity of 99.1% and a specificity of 91.6% on the Aalborg University heart sounds subset of the PhysioNet/CinC 2016 dataset for the task of valve defect recognition. Multiple types of Recurrent Neural Networks (RNN) were used for pathology detection in [17]. The best challenge score of 98.61% on the Physionet/CinC 2016 dataset was obtained using a BiLSTM. Previous studies that used CNN-based, RNN-based, and hybrid methods for PCG classification were reviewed in [18]. It showed that most of the previous works were conducted for abnormal PCG detection. Few of the previous studies were also conducted for heart sound classification into the N, AS, MS, MR, and MVP classes.

The SNMFNet Classifier was proposed in [21] to improve heart sound classification in small samples. It was able to classify heart sounds with 77.6% accuracy on the Physionet/CinC 2016 dataset. A stacked sparse autoencoder deep neural network was used in [22] to classify heart sounds. In this study, time series representations of signals were first converted to time-frequency heat map representations based on fractional Fourier transform-based MFCCs to

obtain an accuracy of 95.5% on the Physionet/CinC 2016 database of heart sounds. A survey of the previous studies has been presented in Table 1.



(a)

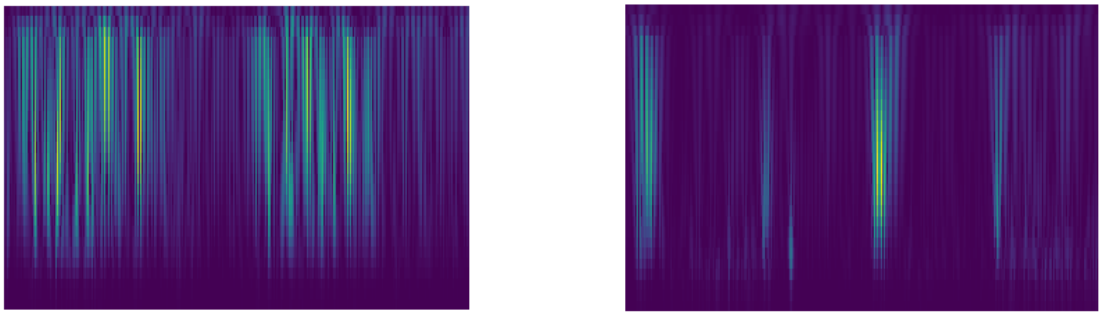


(b)

Figure 2. Presence and absence of murmur in a 1D PCG signal (a) Presence of a murmur (b) Absence of murmur

To the best of our knowledge, at the time of this study, no work has been done for murmur detection on The CirCor DigiScope Phonocardiogram Dataset [23]. This study utilizes signal processing techniques and transfer learning to detect the presence or absence of murmurs in PCG signals recorded from different locations on the chest. The size of the dataset used in this study is

relatively larger (almost double) as compared to the dataset used in previous work [13], thereby enabling more robust classification performance.

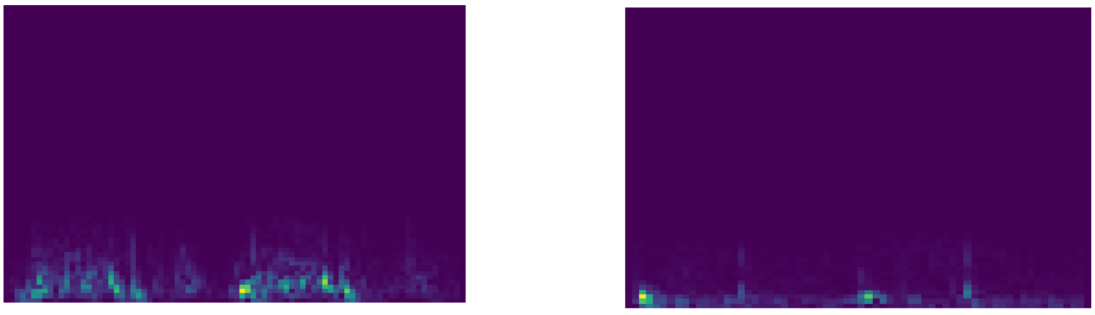


a. Murmur b. No Murmur

Figure 3. Murmur vs No murmur of CWT spectrogram images

a. Murmur b. No Murmur

Figure 4. Murmur vs No murmur of MFCC spectrogram images



a. Murmur b. No Murmur

Figure 5. Murmur vs No murmur of STFT spectrogram images

This study makes the following novel contributions.

(1) It is the first study for murmur detection on the dataset published in [23].

(2) It analyzed the PCG signals for a minimum number of contiguous and complete cycles in order to extract meaningful but least amount of data from PCG signals for spectrogram generation.

(3) It provided the baseline classification accuracy, precision, recall, and F1-score of 89.44%, 92.31%, 74.00%, and 75.51%, respectively.

(4) As compared to the general method of using signals acquired from only one location for training, the proposed model was trained on a dataset of PCG signals acquired from different locations on the chest.

Furthermore, this study makes the following contributions.

(1) It compared multiple transfer learning architectures, namely VGG16, VGG19, ResNet50, and InceptionV3, to detect cardiac murmurs from PCG signals.

(2) It compared different spectrogram techniques, namely CWT, MFCC, and STFT spectrograms, to infer the most relevant one for murmur detection.

(3) It analyzed the application of butterworth and Savitzky-Golay filtering techniques to preprocess PCG signals.

Table 1. Previous works done on heart sound classification

Reference	Features	Classifier	Best Accuracy	Dataset
Almanifi *et al*	Spectrogram	Transfer Learning	87.65%	2011 PASCAL [20]
Demir *et al*	Spectrogram	Transfer Learning+SVM	97.52%	[15]
Fraiwan *et al*	1D time-series	CNN-BiLSTM	87.31%	2016 PhysioNet/CinC [19]
Kucharski *et al*	Time-frequency parameters	DNN	88.20%	2016 PhysioNet/CinC [19]

Latif *et al*	MFCC	DNN	98.86%	2016 PhysioNet/CinC [19]
Abduh *et al*	MFCC	Stacked sparse autoencoder DNN	95.50%	2016 PhysioNet/CinC [19]
Li *et al*	DWT	DNN	92.00%	2016 PhysioNet/CinC [19]
Han *et al*	MFCC map of a segment	SNMFNet Classifier	77.60%	2016 PhysioNet/CinC [19]
Qiang Sun *et al*	Spectrogram	RNN	80.80%	2011 PASCAL [20]

The upcoming sections have been structured as follows. The next section elaborates on the details of the PCG dataset, the experimental setup, and the proposed methodology. In the results and discussions section, multiple performance metrics have been compared for different spectrogram techniques and transfer learning architectures.

2. Methods

2.1. Dataset

The Physionet CirCor Digiscope PCG dataset [23] was used in this study. The data was collected in Northeast Brazil in July-August 2014 and June-July 2015 as part of two mass screening campaigns [24]. Table 2 shows general information regarding the subset of PCG signals from this dataset that were used in this study.

Table 2. General information of PCG signals used from dataset in [23]

Description	Value
Sampling frequency	4000 Hz
Format	.wav
Total number of PCG signals	3163
Number of patients	943
Median length of PCG signals	21.25 seconds

Minimum length of PCG signals	5.15 seconds
Maximum length of PCG signals	62.40 seconds
Number of samples having murmurs	16.38% (499)
Number of samples not having murmurs	78.52% (2391)
Number of samples with missing information of murmur	5.09% (155)

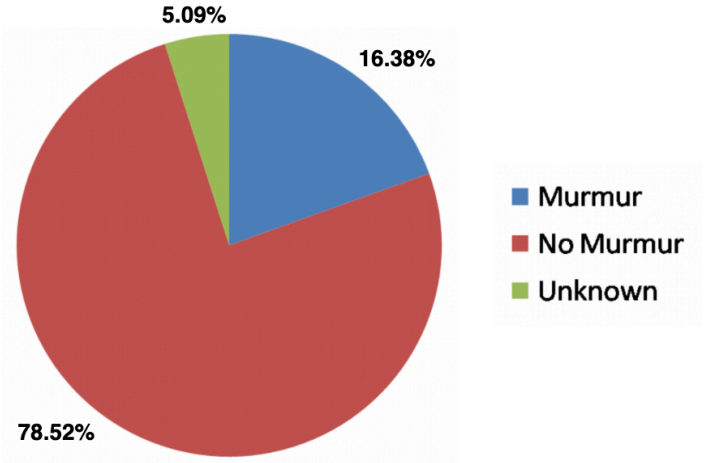


Figure 6. Graphical summary of dataset in terms of diagnostic class

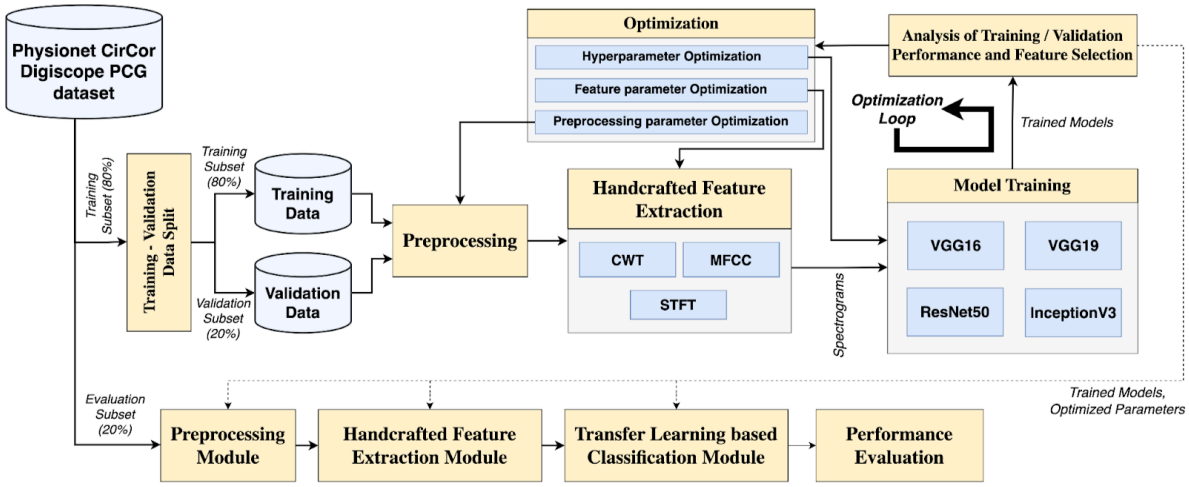


Figure 7. Process to design the proposed system for automated identification of murmurs

Figure 6 shows a graphical summary of the class distribution in the dataset. The dataset is highly imbalanced in favor of the samples having no murmurs (78.52%) as opposed to those with

murmurs (16.38%). There is also a third category that corresponds to those samples for which murmur information is unknown. This class has been disregarded for the purpose of this study since it has no practical application in the task of murmur detection. The class-wise graphical summary of the dataset is shown in figure 6. The dataset was split into the train and test sets in the 80:20 ratio for the purpose of model training and evaluation, respectively.

2.2. Proposed System

The proposed system comprises

(1) a preprocessing module,

(2) a handcrafted feature extraction module, and

(3) a transfer learning architecture.

Detailed descriptions of the system modules are given in the following subsections. The complete process used to create the proposed system is depicted in figure 7.

(1) Preprocessing Module

The preprocessing of raw PCG signals was carried out using a five-step process before the preprocessed output was fed to a 2D CNN architecture. Figure 8 depicts the components of the preprocessing module.

Step 1 - Data preparation

The data preparation process consists of two subtasks - cleaning the data and converting it into a usable format for further processes. Since the dataset contains samples that do not have associated murmur related information, they were removed from consideration in the process of data cleaning.

Step 2 - Segmentation

In order to extract more relevant information from the unsegmented signals, a semantic bound needed to be defined for segmenting the signals, to further use them for spectrogram generation. This bound could then be used to trim the 1D signals. This study used the minimum number of contiguous and complete heart cycles as the bound to trim the PCG signals [25]. It is noteworthy that the general method of segmenting a time-series signal by time duration is not applicable here, since murmurs are more directly associated with cardiac cycles than with any specific time instance in the PCG recording. In order to extract complete cycles from each signal, the complete dataset was analyzed to obtain the minimum number of contiguous complete cycles from each signal. This analysis inferred that each signal contained at least two complete contiguous cycles. Therefore, each signal was trimmed for two cycles in this step. The maximum and minimum time duration of the signals after cycle-based segmentation are shown in Table 3.

Step 3 - Standard scaling

DL models often use gradient-based optimization algorithms like stochastic gradient descent (SGD) to learn optimal weights for the network. These algorithms work better when the input features are standardized and have a similar scale, otherwise the gradient updates may be biased towards features with higher scales, leading to slower convergence and suboptimal results. Additionally, time series signals can have outliers that can affect the performance of the model. Standard scaling helps to reduce the impact of outliers by making the input features more robust to extreme values [26]. Therefore, the PCG signals were preprocessed by centering the signal values around their mean. This was done by subtracting the mean from the sample values and scaling them to unit variance.

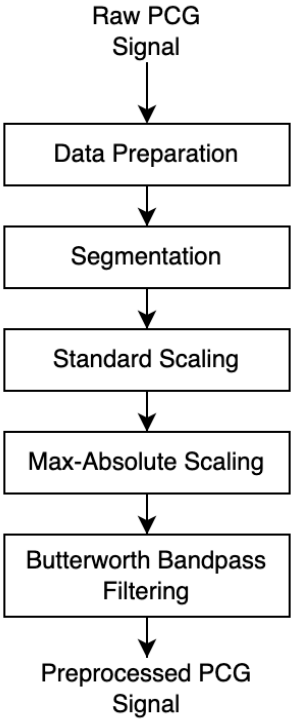


Figure 8. Preprocessing pipeline for PCG signals

Table 3. Time duration of PCG signals after segmentation

	Duration
Minimum length	0.60 seconds
Maximum length	2.64 seconds

Step 4 - Max-absolute scaling

Deep neural networks use activation functions like sigmoid or tanh, which work well with inputs that are normalized to a similar scale between 0 and 1. If the input features are not normalized, the activation functions may saturate or become non-linear, leading to slower convergence and unstable behavior. Since the input PCG signals had a large amount of variation in amplitude values, there was a need to rescale them within the same range (-1 to 1, limits inclusive), thereby making it easier for ML models to learn common patterns [26].

Step 5 - Butterworth bandpass filter

PCG signals comprise various frequency components, such as the low-frequency component generated by cardiac activity and higher frequency components caused by breathing, noise, and other sources. Therefore, a butterworth bandpass filter was employed to eliminate undesired noise and isolate the frequency range relevant to heart sounds [27]. A previous study [28] stated that the frequency range of fundamental heart sounds and murmurs lie between 20-400Hz. Another study [14] utilized a fourth order butterworth bandpass filter with cut-off frequencies of 20-400 Hz to obtain the required frequency ranges from PCG signals. Hence, a fourth order butterworth band pass filter with a cut-off frequency range of 20-400 Hz was used to filter the signals.

(2) Handcrafted Feature Extraction Module

Handcrafted feature extraction techniques are often utilized to extract meaningful features from 1D physiological signals that facilitate further analysis and classification tasks [29]. CWT, STFT, and MFCC are among the commonly used handcrafted feature extraction techniques in 1D signal analysis [30, 31]. CWT separates heart sounds from noise and artifacts by identifying the frequency components of the signals. STFT captures the spectral information of the signal and identifies the time-frequency characteristics of heart sounds. MFCC captures the spectral envelope of the signal and transforms it into a feature set that is suitable for audio processing applications.

In this study, STFT, MFCC, and CWT spectrograms were generated for the preprocessed PCG signals. The parameters used for each transform are listed in Table 4. The time durations of cardiac cycles extracted from different PCG signals were not the same. As a consequence, the

resulting spectrogram images required standardization before being fed as input to deep learning models. Hence, the spectrograms were resized to the size of the input layers (224 X 224) of the subsequent transfer learning models. These spectrograms were further used to train and evaluate transfer learning architectures for the detection of murmurs in PCG signals.

Table 4. Parameters used for CWT, STFT, and MFCC

Transform	Parameter Name	Value
CWT	Mother wavelet used	Mexican hat
	Sampling period for frequencies output	0.00025 seconds
	Method of computation	Discrete linear convolution
STFT	Sampling frequency	4000 Hz
	Number of samples per segment	128
	Number of points to overlap between segments	64
	Window	Hann
	Boundary padding	Zero
	Scaling	Magnitude spectrum
MFCC	Sampling frequency	4000 Hz
	FFT window length	128
	Number of mel bands to generate	16
	Number of samples between successive frames	64
	Exponent of magnitude melspectrogram	1
	Window	Hann
	Boundary padding	Constant

(3) Transfer Learning based Classification Module

Transfer learning techniques were used for the classification of spectrograms of PCG signals in this study instead of custom designed lightweight CNNs due to the following reasons.

a. *Relatively small dataset* - The PCG signals dataset that was used in this study is smaller compared to publicly available datasets of other physiological signals. The principle of transfer learning is generally helpful in dealing with classification when the training dataset is small. Transfer learning models used in this study are ideally expected to have already acquired rich representations of image features, making them capable of recognizing various patterns such as edges and textures that are transferable to PCG classification.

b. *Worked well on handcrafted 2D features* - The generated spectrograms were not generally human-interpretable. Deep transfer learning-based models are more efficient in capturing complex and hierarchical features, and extracting relevant information for classification from the handcrafted 2D dataset without further human involvement.

This study leveraged transfer learning architectures [32], pre-trained on *ImageNet* [33], to establish reasonably accurate classifiers. After an initial observation of training accuracies and survey of recent literature [4] [34] that analyzed multiple transfer learning architectures for spectrogram classification, the following were shortlisted for this study.

a. VGG16
b. VGG19
c. ResNet50
d. InceptionV3

The general details of these architectures are listed in Table 5. The layers added at the classification ends of the architectures used in this study are shown in figure 9.

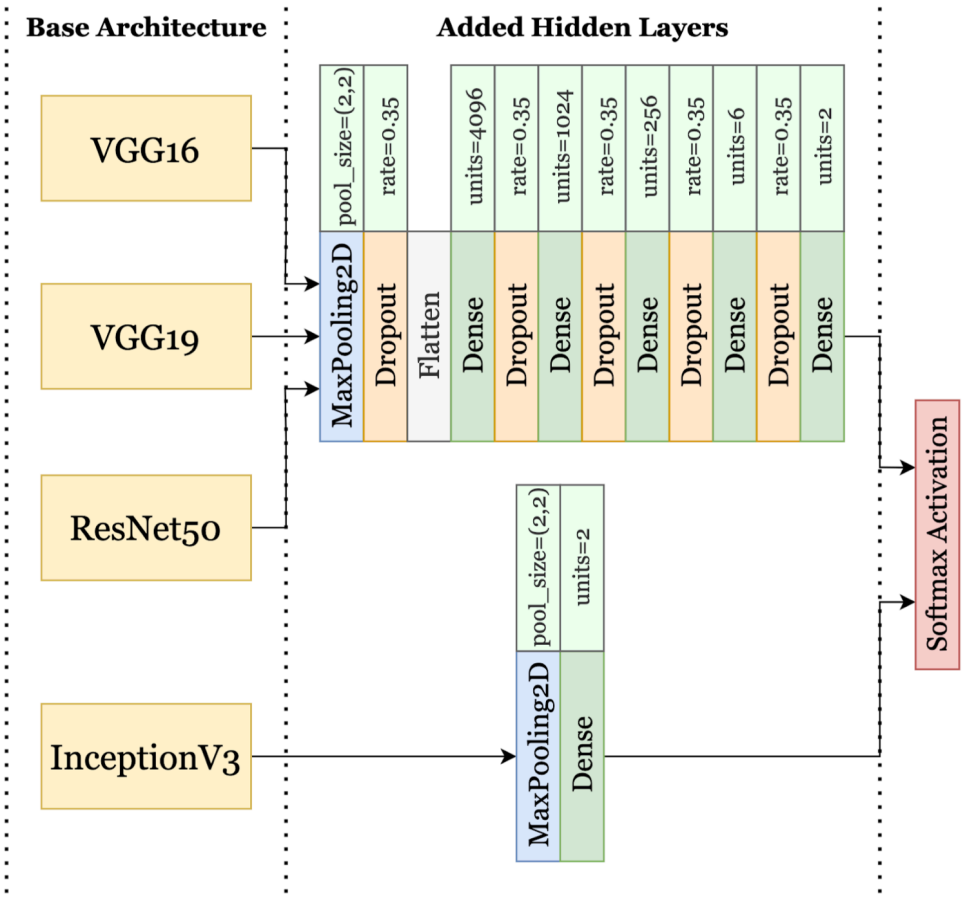


Figure 9. Layers added to the transfer learning architectures

ResNet50 is trained on 26.6 million parameters and has a total depth of 107. It consists of 48 convolutional layers, one 2D max-pooling layer and one average pool layer. VGG16 is trained on 138.4 million parameters and has a depth of 16. The VGG19 model is the same as VGG16 except that it supports 19 layers and is trained on 143.7 million parameters. The InceptionV3 model itself is made up of symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concatenations, dropouts, and fully connected layers. InceptionV3 is trained on 23.9 million parameters and has a depth of 189.

Since initial classification accuracy values with lesser dense and dropout layers at the classification ends of the VGG16, VGG19, and ResNet50 architectures were in the order of ~60-70 %, multiple dense and dropout layers were used. On the contrary, a single dense layer was used at the end of the InceptionV3 architecture since the performance did not improve with additional layers.

CWT, STFT, and MFCC spectrograms capture various time-frequency representations of PCG signals. Using multiple dense and dropout layers enables a model to learn hierarchical features that correspond to different scales and patterns present in these spectrograms. Each dense layer extracts increasingly abstract representations, allowing the model to discern complex relationships between spectrogram elements, potentially leading to improved classification performance. Each dense layer is followed by an activation function, introducing non-linearity to the model. This increased non-linearity allows the network to learn and represent non-linear decision boundaries, which is important when dealing with complex and non-linear relationships between spectrogram features and the presence of murmurs. Additionally, using multiple dense layers offers flexibility during fine-tuning and adaptation of the pre-trained models. The number of units in each dense layer and the depth of the architecture can be adjusted to suit the complexity of the classification task, enabling customization to the specific characteristics of PCG signals.

Dropout layers inserted between the dense layers act as a form of regularization, reducing the risk of overfitting. By randomly dropping out neurons during training, dropout layers help prevent the model from relying too heavily on specific features in the training data and improve its generalization to new, unseen spectrograms.

The final configurations of the transfer learning networks used in this study were determined by carefully selecting the number of dense layers and incorporating multiple dropout layers, to achieve better classification performance while mitigating overfitting. The number of dense and dropout layers and their respective layer-wise parameters were optimized for the highest accuracy values.

Table 5. General details of transfer learning architectures used in this study

Architecture	Parameters	Depth	Top-1 accuracy
VGG16	138.4 million	16	71.3%
VGG19	143.7 million	19	71.3%
ResNet50	26.6 million	107	74.9%
InceptionV3	23.9 million	189	77.9%

2.3. Experimental setup

This section provides details of the performance metrics used to evaluate the proposed system. Additionally, it includes information about the software and hardware configurations utilized for all experiments.

(1) Software setup

Python [35] and Anaconda [36] were employed for programming and creating virtual environments, respectively, for the present investigation. *TensorFlow* [37] and *keras* [38] libraries were utilized for performing deep learning tasks. Library functions from *scikit-learn* [39] and *scipy* [40] were used for scaling and filtering operations. *Pywavelets* [41], *scipy* and *librosa* [42] libraries were used to apply general signal processing operations and generate CWT, STFT, and MFCC spectrograms.

(2) Hardware setup

Table 6 outlines the pertinent hardware specifications of the machine utilized to carry out all experiments.

Table 6. Specifications of hardware used for experiments

Hardware	Specifications
CPU	Model - Intel (R) Xeon (R) Gold 5218 CPU @ 2.30 GHz Number of CPU cores - 64
GPU	Model - Tesla V100-PCIE-32GB Number of GPUs - 2
RAM	Size - 1TB

(3) Performance metrics for evaluation

This study used the metrics of accuracy, recall, precision, F1-score, and ROC-AUC for model selection. Each metric is described in Table 7 in terms of the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Table 7. Performance metrics used in this study

Metric	Description
Accuracy	(TP + TN) / (TP + TN + FP + FN)
Recall	TP / (TP + FN)
Precision	TP / (TP + FP)
F1-score	TP / (TP + 0.5 * (FP + FN))
ROC-AUC	Area under the ROC curve plotted on the FP rate vs TP rate plane

3. Results and Discussion

The results of using different spectrogram techniques on classification and effect of preprocessing techniques are reported in this section. In addition, comparison of accuracy

metrics for pre-trained models with and without butterworth bandpass filter has been made in this section. Furthermore, results of k-fold cross-validation have also been reported.

3.1. Model Hyperparameters

Table 8 shows the final set of hyperparameters that were utilized to configure the training of the models. The training process utilizes the Nadam optimizer, which combines *Nesterov Accelerated Gradient descent (NAG)* and *ADAptive Moment estimation (Adam)*. This optimizer is particularly useful for optimizing neural networks with large amounts of data and parameters.

Table 8. Hyperparameter configuration for training of transfer learning models

S.No.	Name	Values
1	Optimizer Configuration	name = Nadam beta_1 = 0.9 beta_2 = 0.999 epsilon = 1e-07 learning_rate = 0.001
2	Batch Size	32
3	Epochs	100

3.2. Effect of Butterworth Bandpass Filter

The effect of the butterworth bandpass filter on the performance of classification was analyzed for all the tested spectrograms to select the best spectrogram with the filtering process as described in the following subsections.

Comparison of bands of Butterworth Bandpass Filter for classification of spectrograms

On comparing the performance of all feature and model combinations, the CWT and VGG16 combination was found out to be the best. Table 9 shows a comparison of results of using various frequency bands for filtering the signals before generating CWT spectrograms. The results show

that the band of 20-400 Hz is the best filter band for preprocessing signals for VGG16, VGG19, and ResNet50 models, and 500-800 Hz is the preferred filter band for the InceptionV3 model.

3.3. Comparison of performance of all pipelines on the default test dataset

Table 10 shows a comparison of all performance metrics of all pipelines each composed of dataset and model combinations. As shown in Table 10, CWT spectrograms provide the highest classification accuracy on the default test dataset. All the performance metrics for the models trained on CWT spectrograms are shown in figure 10. The confusion matrix for the best performing pipeline (CWT spectrograms + VGG16) is shown in figure 11. The results show that VGG16 turned out to be the most accurate model (89.44%) despite having the least depth as compared to other pretrained models. On the contrary, InceptionV3 performed the worst (76%) despite having the maximum depth.

3.4. Comparison of k-fold cross-validation performance of all pipelines

Table 11 shows a comparison of the k-fold cross-validation performance metrics with k = 5. ResNet50 provided the highest average training accuracy values for CWT, MFCC, and STFT spectrograms. However, the highest average classification accuracy of 89.78% was obtained using CWT spectrograms of the test dataset using the VGG16 model. The cross-validation results reinforce *CWT spectrograms + VGG16* as the best performing pipeline. InceptionV3 performed the worst on all datasets in line with the classification performance on the default test set.

Table 9. Comparison of bands of Butterworth Filter for spectrograms classification

Filter Details	Accuracy (%)			
	VGG16	VGG19	ResNet50	InceptionV3

No Filter	79.00	76.50	76.30	71.50
High-pass, 10 Hz	74.00	76.50	74.00	75.50
Band-pass, 20-200 Hz	77.00	78.50	76.50	78.50
Band-pass, **20-400 Hz**	**89.44**	**87.20**	**85.21**	72.00
Band-pass, 500-800 Hz	77.50	77.00	60.50	**80.00**

Table 10. Performance metrics for all pipelines

Dataset	**Metric**	**Model**			
		VGG16	**VGG19**	**ResNet50**	**InceptionV3**
CWT Spectrograms	Accuracy (%)	**89.44**	87.20	85.81	72.00
	F1-score	59.55	59.34	51.19	**63.16**
	Recall	53.00	**54.00**	43.00	48.00
	Precision	67.95	65.85	63.24	**92.31**
	ROC-AUC	73.88	**74.07**	68.88	72.00
MFCC Spectrograms	Accuracy (%)	**88.06**	85.29	85.12	70.50
	F1-score	56.60	54.05	53.26	**71.22**
	Recall	45.00	50.00	49.00	**73.00**
	Precision	**76.27**	58.82	58.33	69.52
	ROC-AUC	71.04	**71.34**	70.84	70.50
STFT Spectrograms	Accuracy (%)	85.29	85.99	**86.85**	76.00
	F1-score	50.87	47.06	53.66	**75.51**
	Recall	44.00	36.00	44.00	**74.00**
	Precision	60.27	67.92	68.75	**77.08**
	ROC-AUC	68.97	66.22	69.91	**76.00**

Table 11. Results of 5-fold cross-validation

Dataset	**Metric (%)**	**Model**

		VGG16	VGG19	ResNet50	InceptionV3
CWT Spectrograms	Average Training Accuracy	93.99	94.45	**98.63**	90.66
	Average Test Accuracy	**89.78**	89.13	88.40	82.10
MFCC Spectrograms	Average Training Accuracy	94.17	93.70	**98.84**	90.93
	Average Test Accuracy	**88.50**	**88.50**	88.12	87.33
STFT Spectrograms	Average Training Accuracy	92.11	91.01	**98.54**	81.07
	Average Test Accuracy	89.13	88.50	**89.16**	74.73

3.5 Convergence in the training process of best performing models

Figure 12 shows the graphs of accuracy versus epochs during the training phase of the best model with each dataset type. It shows that VGG16, when used with MFCC spectrograms as input, converges the fastest. Loss graphs for the training process of the same models are shown in figure 13. They show that the ResNet50 model trained on STFT spectrograms fluctuates a lot before convergence. On the other hand, the loss values of the VGG16 model fluctuate the least when trained on MFCC spectrograms.

3.6. Training time

The average time taken per epoch to train the transfer learning models with CWT spectrograms is shown in figure 14. Inceptionv3 took the least amount of time (2 seconds), followed by ResNet50 and VGG16 that required 3 seconds per epoch. VGG19 needed the maximum amount of time (4 seconds per epoch) for training.

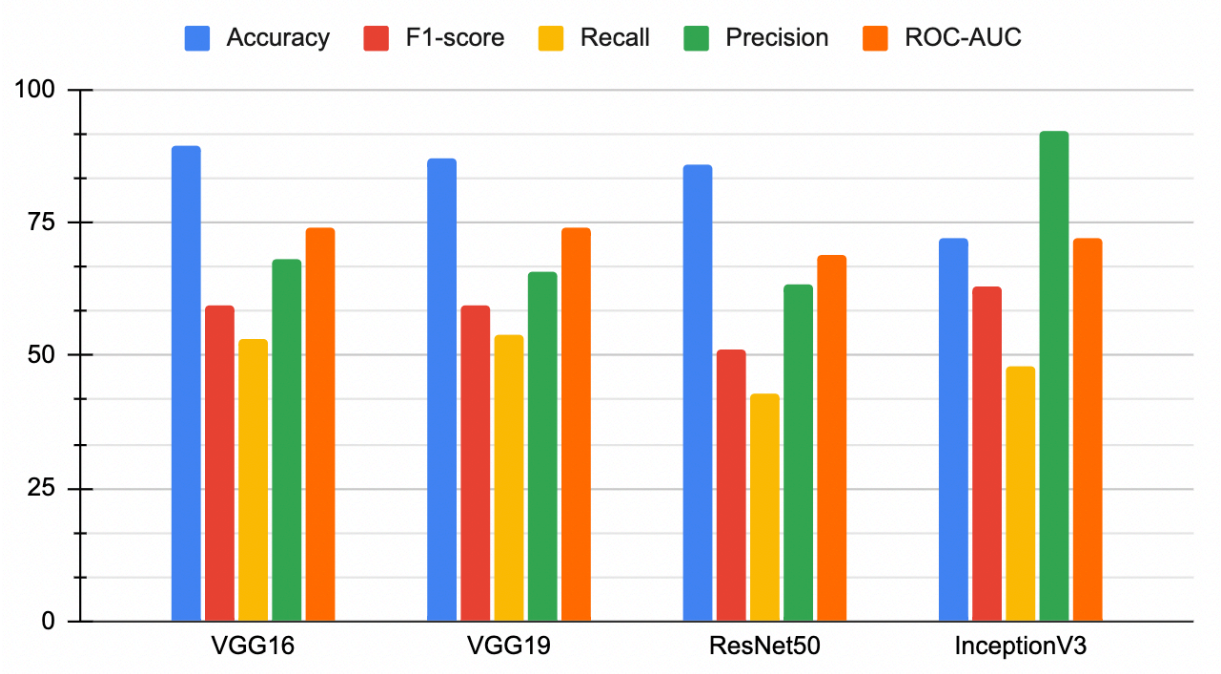


Figure 10. Comparison of performance metrics of all models trained on CWT spectrograms

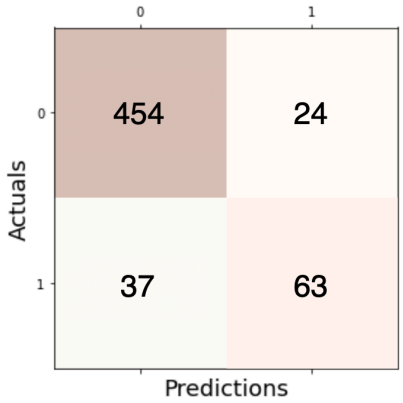


Figure 11. Confusion matrix for best performing pipeline (CWT Spectrograms + VGG16)

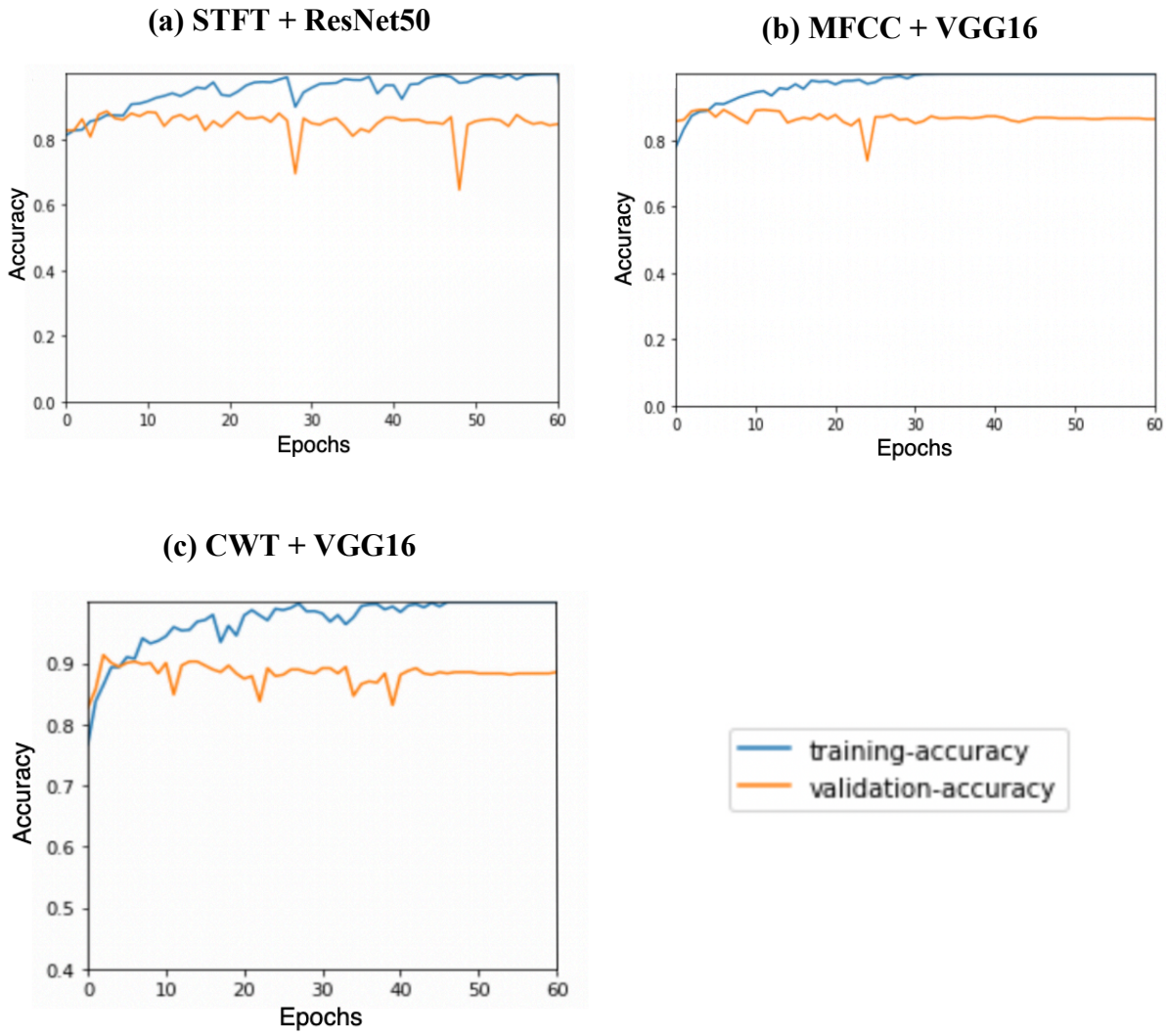


Figure 12. Accuracy and Validation Accuracy vs Epochs for CWT, MFCC and STFT Spectrogram with best performing models

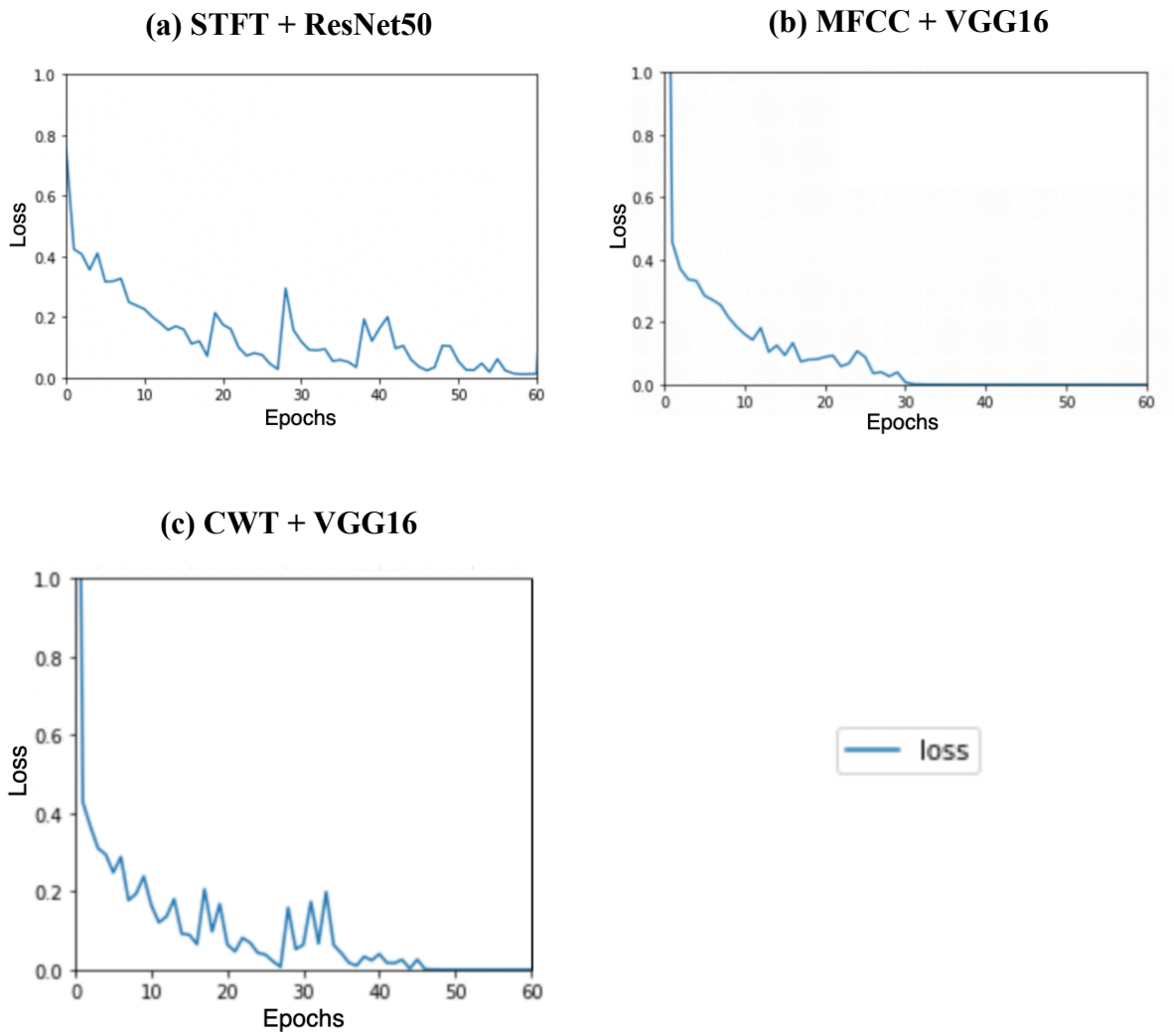


Figure 13. Loss graphs for CWT, MFCC and STFT Spectrograms with best performing models

4. Conclusion

PCG captures heart sound patterns that may potentially contain cardiac murmurs. These murmurs can be an indication of CVDs. However, manual interpretation of PCG signals is challenging due to the presence of external noise and expert human resource requirements. Hence, this study proposed a transfer learning based cardiac murmur detection system for PCG signals.

In this study, filtering and cycle-based segmentation were performed to prepare the signal for the generation of handcrafted features. CWT, MFCC, and STFT spectrograms were the handcrafted features selected for further analyses. VGG16, VGG19, ResNet50, and InceptionV3 were used as the transfer learning architectures for classification with minimal modifications to the ending layers. The final results showed that the most accurate detection of murmurs is possible using the VGG16 model, trained on CWT spectrograms that were generated after signal filtering using a butterworth bandpass filter of band 20-400 Hz.

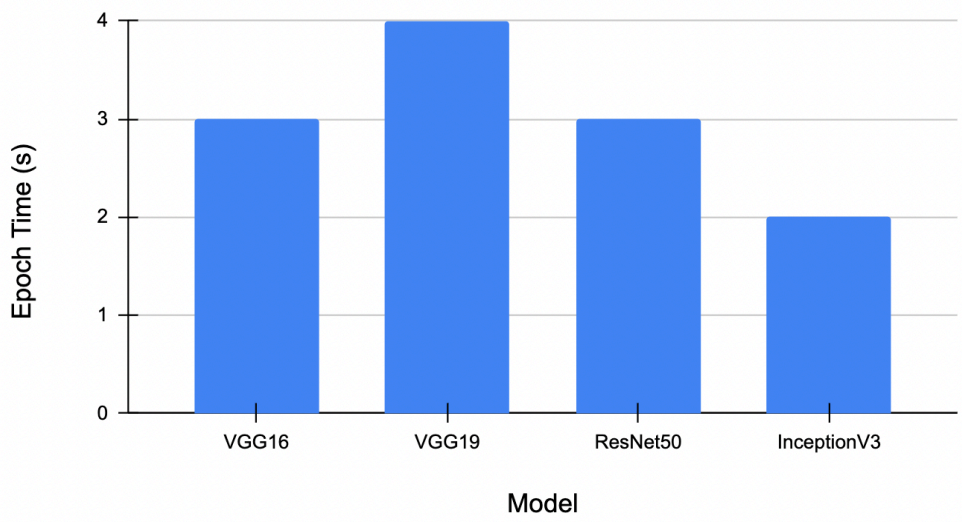


Figure 14. Average time taken per epoch for the training process of the models with CWT spectrograms

At the time of this study, this is the first work on PCG recordings from Physionet's DigiScope CirCor data [23] using transfer learning. The use of pre-trained models greatly reduces the time for training. However, the class distribution in this dataset is highly imbalanced, which limits the usage of the proposed system in real-world setups. The accuracy of the system can be further improved using a larger and balanced dataset that could provide more variation in the data and hence, result in more reliable models.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to express their gratitude towards the Indian Academy of Sciences for providing an excellent research opportunity through their enterprising Summer Research Fellowship Programme. We thank Project Samarth, an initiative of the Ministry of Education (MoE), Government of India, at the University of Delhi South Campus (UDSC), for their support.

References

1. Cardiovascular diseases (CVDs). 2021. [accessed 2023 Apr 18]. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
2. Knuuti J, Wijns W, Saraste A, Capodanno D, Barbato E, Funck-Brentano C, Prescott E, Storey RF, Deaton C, Cuisset T, et al. 2019. 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes: The Task Force for the diagnosis and management of chronic coronary syndromes of the European Society of Cardiology (ESC). Eur Heart J. 41(3):407–477.
3. Artinian NT, Fletcher GF, Dariush Mozaffarian, Kris-Etherton P, Linda Van Horn, Lichtenstein AH, Shiriki Kumanyika, Kraus WE, Fleg JL, Redeker NS, et al. 2010. Interventions to promote physical activity and dietary lifestyle changes for cardiovascular risk factor reduction in adults. Circulation. 122:406–441.
4. Vikhe P, Nehe N, Ahmednagar L, Thool V. 2009. Heart sound abnormality detection using short time fourier transform and continuous wavelet transform. Proceedings of the

Second International Conference on Emerging Trends in Engineering & Technology, p. 50–54

5. Chen J, Guo Z, Xu X, Zhang L, Teng Y, Chen Y, Woźniak M, Wang W. 2023. A robust deep learning framework based on spectrograms for heart sound classification. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 1–12.
6. Qin L. 2023. A review of heart sound and its research methods. Academic Journal of Computing & Information Science. 6(3):84–91.
7. Yildirim M. 2022. Automatic classification and diagnosis of heart valve diseases using heart sounds with MFCC and proposed deep model. Concurrency and Computation: Practice and Experience [Internet]. 34(24):e7232.
8. Rashid N, Saha S, Mohseu Rashid Subah, Rizwan Ahmed Robin, Syed, Ahmed S, Talha Ibn Mahmud. 2022. Heart abnormality detection from heart sound signals using MFCC feature and dual stream attention based network. arXiv [cs.SD].
9. Bao X, Xu Y, Kamavuako EN. 2022. The effect of signal duration on the classification of heart sounds: A deep learning approach. Sensors. 22(6):2261.
10. Shuvo SB, Alam SS, Ayman SU, Chakma A, Barua PD, Acharya UR. 2023. NRC-Net: Automated noise robust cardio net for detecting valvular cardiac diseases using optimum transformation method with heart sound signals. Biomedical Signal Processing and Control. 86:105272.
11. Bao X, Xu Y, Lam H-K, Trabelsi M, Chihi I, Sidhom L, Kamavuako EN. 2023. Time-Frequency distributions of heart sound signals: A Comparative study using convolutional neural networks. Biomedical Engineering Advances [Internet]. 5:100093.

12. Petrolis R, Paukstaitiene R, Rudokaite G, Macas A, Grigaliunas A, Krisciukaitis A. 2022. Convolutional neural network approach for heart murmur sound detection in auscultation signals using wavelet transform based features. In: 2022 Computing in Cardiology (CinC). vol. 498. p. 1–4.
13. Rashed O, Fakhri A, Mohd, Rabiu Muazu Musa, Majeed A. 2022. Heartbeat murmurs detection in phonocardiogram recordings via transfer learning. Alexandria Engineering Journal. 61(12):10995–11002.
14. Khan KN, Khan FA, Abid A, Olmez T, Dokur Z, Khandakar A, Chowdhury MEH, Khan MS. 2021. Deep learning based classification of unsegmented phonocardiogram spectrograms leveraging transfer learning. Physiological Measurement. 42(9):095003.
15. Alkhodari M, Fraiwan L. 2021. Convolutional and recurrent neural networks for the detection of valvular heart diseases in phonocardiogram recordings. Computer Methods and Programs in Biomedicine. 200:105940.
16. Kucharski D, Grochala D, Kajor M, Kańtoch E. 2018. A deep learning approach for valve defect recognition in heart acoustic signal. In: Borzemski L, Świątek J, Wilimowska Z, editors. ISAT 2017. Proceedings of 38th International Conference on Information Systems Architecture and Technology: Springer International Publishing. p. 3–14.
17. Latif S, Usman M, Rana R, Qadir J. 2018. Phonocardiographic sensing using deep learning for abnormal heartbeat detection. IEEE Sensors Journal. 18(22):9393–9400.
18. Chen W, Sun Q, Chen X, Xie G, Wu H, Xu C. 2021. Deep learning methods for heart sounds classification: A systematic review. Entropy. 23(6):667.

19. Liu C, Springer D, Li Q, Moody B, Ricardo Abad Juan, Chorro FJ, Castells F, José Millet Roig, Silva I, Johnson AEW, et al. 2016. An open access database for the evaluation of heart sound algorithms. Physiological Measurement. 37(9):2181.
20. Bentley P, Nordehn G, Coimbra M, Mannor S. 2011. The PASCAL classifying heart sounds challenge 2011 (CHSC2011) Results. [accessed 2023 Aug 30]. http://www.peterjbentley.com/heartchallenge/index.html.
21. Han W, Xie S, Yang Z, Zhou S, Huang H. 2019. Heart sound classification using the SNMFNet classifier. Physiological Measurement. 40(10):105003.
22. Abduh Z, Nehary EA, Wahed MA, Kadah YM. 2019. Classification of heart sounds using fractional fourier transform based mel-frequency spectral coefficients and stacked autoencoder deep neural network. Journal of Medical Imaging and Health Informatics. 9(1):1–8.
23. Oliveira J, Renna F, Costa P, Nogueira M, Oliveira AC, Elola A, Ferreira C, Jorge A, Bahrami Rad A, Sameni R, et al. 2022. The CirCor DigiScope Phonocardiogram Dataset. PhysioNet. [accessed: 2023 Aug 30]. https://physionet.org/content/circor-heart-sound/1.0.0/.
24. Oliveira J, Renna F, Costa PD, Nogueira M, Oliveira C, Ferreira C, Jorge A, Mattos S, Hatem T, Tavares T, et al. 2022. The CirCor DigiScope dataset: From murmur detection to murmur classification. IEEE Journal of Biomedical and Health Informatics. 26:2524–2535.
25. Khan FA, Abid A, Khan MS. 2020. Automatic heart sound classification from segmented/unsegmented phonocardiogram signals using time and frequency features. Physiological Measurement. 41(5):055006.

26. Shukla U, Saxena GJ, Kumar M, Bafila AS, Pundir A, Singh S. 2021. An improved decision support system for identification of abnormal EEG signals using a 1D convolutional neural network and savitzky-golay filtering. IEEE Access. 9:163492-163503.
27. Singh SA, Meitei TG, Majumder S. 2020. 6 - Short PCG classification based on deep learning. In: Agarwal B, Balas VE, Jain LC, Poonia RC, Manisha, editors. Academic Press. p. 141–164.
28. Daliman S, Sha'ameri AZ. 2003. Time-frequency analysis of heart sounds and murmurs. Proceedings of Fourth International Conference on Information, Communications and Signal Processing, and the Fourth Pacific Rim Conference on Multimedia. Vol. 2, p. 840–843.
29. Cheng X, Huang J, Li Y, Gui G. 2019. Design and application of a laconic heart sound neural network. IEEE Access. 7:124417-124425.
30. Ismail S, Ismail B. 2023. PCG signal classification using a hybrid multi round transfer learning classifier. Biocybernetics and Biomedical Engineering. 43(1):313–334.
31. Quiceno-Manrique AF, Godino-Llorente JI, Blanco-Velasco M, Castellanos-Dominguez G. 2010. Selection of dynamic features based on time--frequency representations for heart murmur detection from phonocardiographic signals. Annals of Biomedical Engineering. 38(1):118–137.
32. Roy TS, Roy JK, Mandal N. 2022. Classifier identification using deep learning and machine learning algorithms for the detection of valvular heart diseases. Biomedical Engineering Advances. 3:00035.

33. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009. Imagenet: A large-scale hierarchical image database. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. p. 248–255.
34. Chen X, Guo X, Zheng Y, Lv C. 2023. Heart function grading evaluation based on heart sounds and convolutional neural networks. Physical and Engineering Sciences in Medicine. 46(1):279–288.
35. Van Rossum G, Drake FL. 2009. Python 3 Reference Manual. CA. CreateSpace.
36. Anaconda Software Distribution, Anaconda Documentation. 2020. Anaconda Inc.
37. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al. 2016. TensorFlow: A system for large-scale machine learning. p. 265–283.
38. Chollet F. 2015. Keras. [accessed 2023 August 30]. https://github.com/fchollet/keras.
39. Pedregosa F, Gaël Varoquaux, Alexandre Gramfort, Michel V, Thirion B, Grisel O, Mathieu Blondel, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine learning in python. Journal of Machine Learning Research. 12:2825–2830.
40. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 17(3):261–272.
41. Lee G, Gommers R, Waselewski F, Wohlfahrt K, O'Leary A. 2019. PyWavelets: A Python package for wavelet analysis. Journal of Open Source Software. 4(36):1237.
42. McFee B, Raffel C, Liang D, Daniel P.W. Ellis, McVicar M, Battenberg E, Nieto O. 2015. librosa: Audio and Music Signal Analysis in Python. In: Huff K, Bergstra J, editors. Proceedings of the 14th Python in Science Conference. Vol. 8. p. 18–24.

Social Network Analysis and Mining (2023) 13:141
https://doi.org/10.1007/s13278-023-01145-1

ORIGINAL ARTICLE

Identifying discernible indications of psychological well-being using ML: explainable AI in reddit social media interactions

Pahalage Dona Thushari[1] · Nitisha Aggarwal[2] · Vajratiya Vajrobol[2] · Geetika Jain Saxena[3] · Sanjeev Singh[2] · Amit Pundir[3]

Received: 9 May 2023 / Revised: 25 September 2023 / Accepted: 26 September 2023

Abstract

Psychological well-being is a multidimensional construct and identifying it using a systematic, comprehensive approach offers insights fundamental to critical outcomes. Social networks are valuable resources for research, providing a pragmatic way of generating empirical evidence on psychological well-being based on the textual indicators across different populations. This study analyzed the information on various Reddit social media groups dedicated to mental health. The classes, namely depression, anxiety, bipolar, and SuicideWatch, using the SWMH dataset have been analyzed. The text-based interactions of persons with mental illness have common motifs like negative language and expressions like 'hopelessness,' 'emptiness,' or 'helplessness.' Topic modeling identified recurring themes and subjects that helped classify discernible factors influencing mental health. Classifiers for multiclass classification to classify targeted mental health issues based on users' network behavior and posts were trained and tested to get predictions on context (e.g., MentalBERT) and non-context-based (e.g., LR and NB) models. The MentalBERT model outperformed the other eight baseline models with an average accuracy of 76.70%, which is 4% more than reported in previous studies. Explainable AI was used to examine the trustworthiness of each model, and the explanations were evaluated using the LIME model. Explainability is crucial as mental health data characterizes syndromes, outcomes, disorders, and signs/symptoms exhibiting probabilistic interrelationships with each other. Explanations of these intricate interconnections can assist the extensive research around the model of well-being and interventions intended to improve the human condition and distill positive human functioning.

Keywords Topic modeling · BERTopic · Mental health · Reddit · LIME · Multiclass classification

✉ Amit Pundir
amitpundir@mac.du.ac.in

Pahalage Dona Thushari
thusharipahalage@gmail.com

Nitisha Aggarwal
nitisha@south.du.ac.in

Vajratiya Vajrobol
tiya101@south.du.ac.in

Geetika Jain Saxena
gsaxena@mac.du.ac.in

Sanjeev Singh
sanjeev@south.du.ac.in

[1] Department of Software Engineering, Delhi Technological University, New Delhi, Delhi 110042, India

[2] Institute of Informatics and Communication, University of Delhi, New Delhi, Delhi 110021, India

[3] Department of Electronics, Maharaja Agrasen College, University of Delhi, New Delhi, Delhi 110096, India

1 Introduction

The condition of one's mental health significantly impacts the person suffering and the community (Benrouba and Boudour 2023; Ji et al. 2022). According to the United Nations report, in 2022, one in eight people globally suffers from a mental health disorder (World mental health report: Transforming mental health for all - executive summary 2022a). The report has underscored the disparities in mental healthcare, underscoring the global prevalence of high mental health needs alongside frequently inadequate and insufficient responses. Furthermore, it has proposed measures to enhance mental healthcare systems, rendering them more accessible and effective for everyone. The delayed diagnosis of mental illnesses, resulting in delayed preventive interventions, has contributed to approximately 703,000 annual suicide-related deaths across various age groups, genders, and geographical regions. Presently, suicide stands as a leading

Published online: 25 October 2023

cause of mortality, surpassing other factors like malaria, AIDS, breast cancer, war, and homicide on a global scale. To identify high-risk demographics, nations should gather and scrutinize disaggregated data encompassing variables such as gender, age, and suicide methods. These data hold paramount importance in grasping the scope of the issue and facilitate the tailoring of interventions to suit the unique requirements of at-risk populations while adapting to evolving trends (Suicide data: Mental Health and Substance Use 2021). The diagnostic and statistical manual of mental disorders (DSM-V), published by the American Psychiatric Association, defines a mental disorder as a set of symptoms that cause significant disruption in a person's thinking, emotional regulation, or mental functioning that is consistent with psychological, biological, or developmental processes. Mental illnesses are typically accompanied by severe suffering or impairment in vital social, occupational, or other tasks (Stein et al. 2021). Unfortunately, more than 70% of individuals with mental disorders worldwide lack primary care and treatment due to limited resource environments (Kilbourne et al. 2018).

Mental health professionals study feelings, thoughts, behavior patterns, and other tests to diagnose illness. The availability of such resources is a critical constraint for conducting these studies, as health records of patients' emotional and intimate feelings are not shared due to privacy and confidentiality concerns. Data scarcity, therefore, restricts the research in this domain and severely limits designing accurate methods and techniques for diagnosing mental illness (Hanna et al. 2018). Moreover, most mental disorders have similar characteristics, making distinguishing between mental health diseases for diagnostic purposes difficult. In addition, the availability of mental health experts per million population, particularly in Low- and middle-income countries (LMICs), is deficient (Wainberg et al. 2017). Handling of few mental illnesses, such as depression, anxiety, and bipolar disorder, is challenging as they do not have cures. However, behavior therapy and medication can significantly control the illness and the quality of patients' lives may significantly improve. Therefore, reducing stress in peoples' lives is crucial by knowing the cause of illness and minimizing its impact on their daily lives.

Social network platforms allow individuals apprehensive about face-to-face interactions to share their thoughts and feelings with a larger community. These social media platforms are easy to access and do not discriminate based on age, gender, socio-economic status, race, or ethnicity. Communicating with a larger community of people suffering from similar concerns in online forums helps them regulate their emotions. Their interactions, in the form of texts, become an invaluable resource facilitating analysis of textual signs of psychological health problems (Dao et al. 2015; Kamarudin et al. 2021). Among the popular social media networks, the Reddit platform is preferred by young individuals for sharing their emotions as it provides anonymity and a handy platform for discussions on mental health issues (Boettcher 2021). Researchers using such data can analyze the emotions expressed to understand the mental issues the users suffer and design tools to find the types, stages and cost-effective solutions (Ren et al. 2021).

Recent years witnessed an abrupt growth in the analysis of social media data to investigate a variety of health issues ranging from the effects of allergies and Covid-19 (Huang et al. 2022; Zhou et al. 2021; Kathy et al. 2015) and sentimental analysis on Covid-19 vaccine (Alotaibi et al. 2023; Verma et al. 2023) to emotions and mental health conditions (Saha et al. 2016; Kim et al. 2020; Lin et al. 2023). Sentiment analysis is applied to identify the underlying sentiment or emotion expressed in a piece of text. The sentiment of a person's post, certain language and linguistic patterns can provide insights into their mental state. Natural language processing (NLP) analyzes such data and allows for diagnosing disorders and developing treatment strategies. Numerous methods have been developed to map the semantic relations between textual data (Qi and Shabrina 2023; Rizvi et al. 2023). Deep learning (DL) models recurrent neural network (RNN), long short-term memory (LSTM), transformer, convolutional neural networks (CNN), bidirectional encoder representations from transformers (BERT) and other hierarchical attention networks (HAN) have been applied to analyze text at different levels, such as document level and sentence level, for identifying mental health-related concerns. BERT and other transformer-based models employ attention-based mechanisms; hence, they offer several benefits over traditional models like CNN and LSTM in the context of NLP. This attention mechanism helps identify significant words, phrases, or patterns that contribute to the overall meaning or sentiment as they focus on different parts of the input text. They also consider the neighboring words and their relationships to allow for a better understanding of the meaning and semantics of a word within a sentence to capture contextual information effectively. Traditional models like CNN and LSTM process data sequentially and hence are limited in capturing long-range dependencies in text, while transformer models can capture dependencies between words regardless of their positional distance through parallel processing. Additionally, contextual methods, like the MentalBERT model, consider the context and relationships within the text to understand its meaning. They capture nuances and dependencies, making them suitable for mental health discussions. Non-contextual methods like Bag of Words and TF-IDF analyze words in isolation without considering the context. They extract features using statistical or rule-based techniques, providing insights without contextual nuances. Non-contextual methods may not capture the subtleties and dependencies present in the text as effectively as

contextual methods. However, they can still provide valuable insights, especially when the specific context is not crucial for the analysis. This paper uses recent developments in NLP to find the common underlying topics in communities with mental health issues using their social media interactions. Our work has three focuses of interest:

1. The extraction and evaluation of significant themes and subjects for sensitive mental health issues using topic modeling;
2. Comparative analysis of context-based and non-context-based machine learning (ML) classifiers to automate the classification process of textual data on mental health issues, and
3. Using model interpretability, analyze the secular associations present in the distinctive attributes.

Identifying the variables contributing to mental health issues based on social media interactions can reveal common factors. These factors may subsequently lead to determining new perspectives of interventions and strategies to achieve better mental health solutions. This research uncovered the main themes and patterns within mental health-related content derived from a social media platform, employing the topic modeling technique. Through these identified themes, we gain access to the underlying semantic structure embedded within the text. For instance, recurrent themes in the study encompassed topics like exams, school, college, friends, and relationships, highlighting their interconnectedness with mental health. Moreover, the outcomes revealed that context-driven models such as BERT and MentalBERT exhibited superior post-classification accuracy compared to traditional models like logistic regression (LR) and random forest (RF). To gain insights into the classification model results, we employed LIME, which indicated that BERT and MentalBERT classified posts based on contextual factors. In addition to this, our findings indicated that individuals with mental health concerns may potentially exhibit a likelihood of multiple mental disorders, a facet elucidated through the predictive capability of LIME explanations. The rest of the paper has been arranged as follows. Section 2 reports the related work, while the methodology is shown in Sect. 3. Section is about the findings of the study. Section 4 is the discussion related to the results obtained and their significance. Finally, in Sect. 5, the conclusions and future work are discussed.

2 Related work

Recent advancements in NLP and deep learning have positively influenced the analysis of social networking media interactions of communities with mental health issues. Social media data on platforms such as Twitter and Reddit (Liu et al. 2022) are of prime interest to researchers as they are real-time, readily available and help to reach a wider audience at a low cost. The research by Ji et al. (2022) describes a relation network with an attention mechanism to identify mental disorders and suicidal ideation with associated risk indicators. The SWMH (SuicideWatch and Mental Health) real-world dataset contains subreddit data divided into multiple classes, which we used in this study. They enhanced text representation and measured sentiment score and latent topics by lexicons. Their best-performing model for SWMH data achieved 64.74% accuracy for relation networks. For the same dataset, another study introduced pre-trained language models, namely MentalBERT and MentalRoBERTa (Ji et al. 2021) and achieved 72.16% best F1-score for the MentalRoBERTa model. These models were trained on domain-specific language for mental healthcare and are publicly available.

The work by Guntuku et al. (2017) investigated potential ways to use screening surveys on social media to predict mental health disorders. They detected symptoms associated with mental illness from Twitter, Facebook, and web forums and suggested that AI-driven methods can detect mental illnesses. An earlier study by Gemmell et al. (2019) investigated how to automatically recognize informal patterns in the language retrieved from online forums for borderline personality disorder patients as well as bipolar disorder patients. The top 10 phrases and terms were found to best describe each cluster using k-means clustering on the Reddit data. Another study by Kotenko et al. (2021) on the evaluation of the Mental Health of Social Network Users (Pushshift Reddit Dataset) calculated the emotion lexicon, used Latent Dirichlet Allocation (LDA) for topic modeling and classification using the fastText classifier and achieved 96% F1-score. Traditional approaches, such as LDA (Blei et al. 2003), failed to capture the semantic relationships and were not domain-specific, providing subjects irrelevant to the context.

Researchers widely prefer DL techniques in analyzing social media data to detect mental disorders. A study by Gkotsis et al. (2017) classified mental health conditions using feedforward and convolutional neural networks (CNN) based on the Reddit dataset. This Reddit dataset is further grouped through a semi-supervised technique to create subreddits of 11 mental health-related themes. The best-performing CNN model showed 71.37% accuracy in their study. Another study by Islam et al. (2018) on Facebook data detected depression using psycholinguistic measurements and ML algorithms, including the decision tree (DT), which outperformed all other techniques. Zanwar et al. (2022b) conducted a multiclass classification using hybrid and ensemble transformer models on the self-reported mental health diagnoses (SMHD) dataset and the Dreaddit dataset

using BERT, RoBERTa, and bidirectional long short-term memory (BiLSTM), achieving a macro F1-score of 31.40% across 5 folds.

Interpretability of models becomes crucial as classification alone is insufficient for sensitive applications such as mental health analysis. Explainable AI has emerged as an effective technique to address this problem (Gunning et al. 2019). One such technique, the local interpretable model-agnostic explanations (LIME) (Ribeiro et al. 2016), provides a factual implementation to explain respective predictions. This paper applied the LIME algorithm to present a trustable explanation. Hu et al. conducted a multiclass classification on Covid-19 related mental health data using the post hoc system and ante-hoc method to analyze and explain the factors that impact mental health during the pandemic (Hu and Sokolova 2021). Another recent study (Saxena et al. 2022) on CAMS (Garg et al. 2022) dataset used LIME and integrated gradient (IG) methods to find explanations for reasons related to inconsistency in the accuracy of multiclass classification. Our study found meaningful topics within mental health discussions to gain insights and interpret the findings of individuals suffering from mental health illness using a novel technique, BERTopic (Grootendorst 2022c). We analyzed the possible occurrences of topics for four clinically identified mental health issues, including bipolar disorder, anxiety, suicidal thoughts, and depression. The study has been extended using classification techniques to classify the text corpus into each category. Post classification, the trustworthiness of each model's prediction ability was investigated using local explanations for a given instance using explainable AI.

3 Methodology

3.1 Dataset

Given its novelty and accessibility, we have examined the Reddit SWMH (Ji et al. 2022) dataset. The dataset comprised texts from mental health-related reddit subreddits of anxiety, depression, bipolar, and suicidewatch, a total of 54,412 texts in five classes. This dataset is anonymous; hence, the absence of any identifiable details regarding the individuals in the dataset ensures that the privacy and confidentiality of the participants remain uncompromised. The SWMH dataset's purpose is to identify associated mental health conditions-related variables. This study removed the 'self.offmychest' class and duplicate texts. This class was removed as it did not indicate any direct relation to mental health illnesses, such as depression, anxiety, bipolar disorder, and suicidal thoughts. Our study examined 46,103 instances from the four classes, namely self.Anxiety, self.bipolar, self.depression, self.SuicideWatch. This dataset was analyzed to find themes and patterns of suicidality and mental illnesses.

The total number of texts evaluated for each dataset is shown in Fig. 1. We assessed and contrasted the data available on social media for various mental disorders with recurring themes and patterns associated with mental illnesses. The text samples in Table 1 represent each category.

Word clouds were created to visualize the importance of words in the context. The word cloud generated and shown in Fig. 2, using the text corpus, reveals the frequently used words "feel," "life," and "one" to convey emotions. Words such as 'anxiety,' 'depression,' and 'people' tend to stand out since the users have been sharing texts with phrases like "*I feel anxious around people*" and "*Is there anyone to talk about depression?*" for example. Social media users frequently utilize social platforms allowing for more open discussions about delicate subjects like mental health (Yazdavar et al. 2018).

3.2 Process architecture

The process architecture of this experiment consists of preprocessing texts, topic modeling, classification and local explanations, as shown in Fig. 3. The pre-processing pipeline was further segmented into noise removal and normalization. As for noise removal, the datasets were processed to remove duplicate text, links, text in square brackets, terms with numbers, and lowercase text conversion. In addition, stop words and largely standard English terms lacking crucial information and meaningless and misspelled words were removed. The NLTK library (natural language toolkit) was initially used for this purpose, and the tokenization step was then completed. Tokenization is a crucial process of dividing a text into token-sized pieces allowing for interpreting the text's meaning via word analysis. Several library functions, such as sci-kit-learn countvectorizer for ML models and Tensorflow Tokenizer for CNN and LSTM models, were

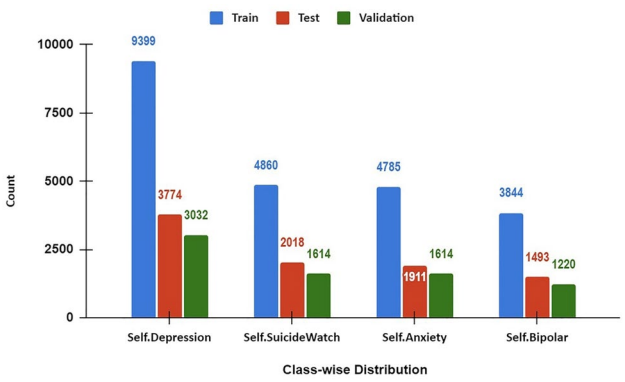


Fig. 1 Class-wise distribution of processed training, testing, and validation sets

Table 1 Samples of text from each class

Label	Text sample
Anxiety	Why do my arms, and muscles jerk when I feel anxious? When my anxiety is particularly bad, I find that my muscle control is horrible. When I try to grip something, my muscles will simply twitch. It feels and appears incredibly strange. Even turning my head in one direction causes it to jerk. I really hate this. It gives me the impression that there is a problem with me
Bipolar	Having recovered from a manic episode and considering scheduling a meeting? Since my fiancé broke up with me two months ago, I've apparently been experiencing manic episodes, and now I'm beginning to collapse
Depression	Right I must speak with someone. Every time I try to talk to my friends about my problems, they either tell me I'm acting silly or that things don't really matter, that nobody cares, or that I'm just being a little bitch. They could be correct, but those things have been bugging me for about two to three months now, and I really need to get it off my chest because it's getting exhausting. Although I frequently consider suicide, I lack the courage to carry it out
SuicideWatch	After a few days of abstinence, I cut today. Even though it wasn't much time, I felt pretty proud of myself for not doing it. And right now, slashing myself and taking my own life is all I can think about. I should have done it a week ago, but I decided not to, which was a terrible error. I overdosed four times in the last five months, and I wish one of them had killed me. I know if I leap off the nearby overpass of the highway, I'll get better results. This world is not suited for me. I have no place here. I just want to fall asleep and stay asleep forever

Fig. 2 Word cloud for training data to visualize important words in the context of mental health

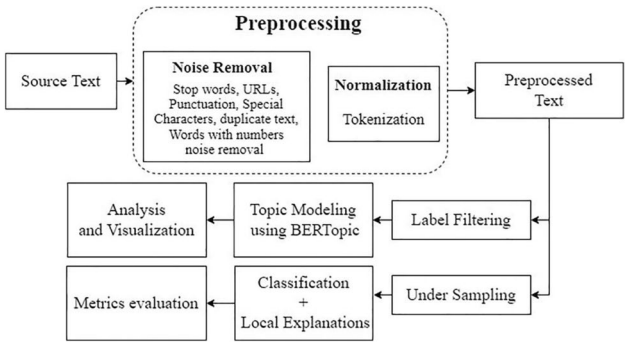


Fig. 3 Architecture of the approach used in this study to identify mental health concerns

used to complete this step. The Transformer models were used with their respective tokenizers provided by the HuggingFace platform. Normalization, a pre-processing step used to enhance the text's quality and prepare it for machine processing, was done. While conducting the classification, undersampling and class weights were applied to deal with the class imbalance.

3.2.1 Topic modeling

By examining the terms in the source texts, topic modeling seeks to identify the themes that permeate a corpus. Topic modeling is an unsupervised ML method of identifying

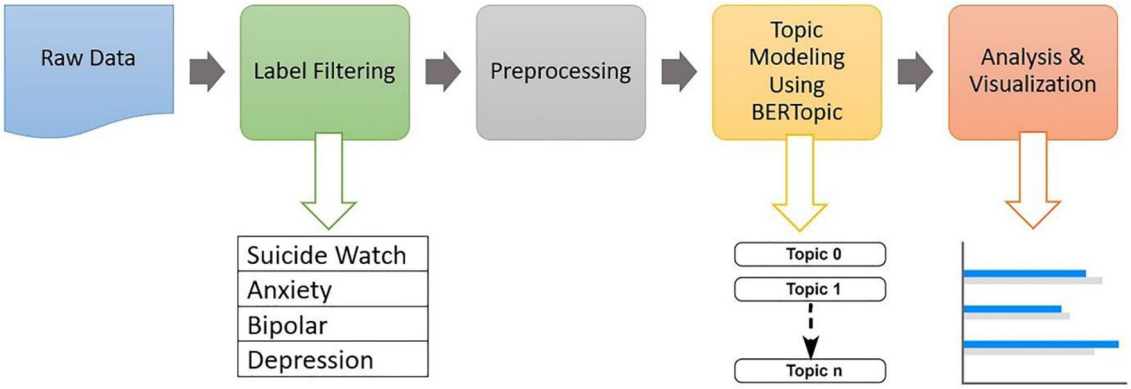


Fig. 4 Architecture of the approach used in this study for topic modeling

or extracting topics by spotting patterns, much like clustering algorithms that separate data into various sections. Transformers, C-TF-IDF and BERTopic, use dense clusters to generate simple to understand topics while preserving critical terms from the topic descriptions. Earlier BERTopic-based studies (Sangaraju et al. 2022; Abuzayed and Al-Khalifa 2021) found that it is adaptable and offers distinct topics compared to LDA. BERTopic has expanded the classical cluster embedding approach by utilizing cutting-edge language models and a class-based TF-IDF(C-TF-IDF) process to create topic representations. The model uses three phases to generate topic representations. Each document is first transformed using a trained language model into its embedding representation. The dimensionality of the generated embeddings is then decreased before clustering them to improve the clustering procedure. Due to the stochastic nature of the UMAP method used for dimensionality reduction while clustering, the model gives different results for training instances. Finally, topic representations are retrieved from the document clusters using a customized class-based form of TF-IDF. The model is significantly more flexible and easier to use because the grouping of documents and generating topic representations is carried out independently (Grootendorst 2022c). C-TF-IDF is utilized in the third stage to extract essential terms at each cluster on a class-based term frequency or inverse document frequency. In BERTopic, TF-IDF was modified to operate on a cluster/category/topic level rather than at the document level to accurately represent the subjects from the bag-of-words matrix. This modified TF-IDF representation, known as C-TF-IDF, considers the differences between documents in each cluster by treating each cluster as a single document rather than a collection of documents. The frequency of the word x in class c, where c is the cluster that we previously established, is extracted. The class-based TF representation is the outcome of this. L1-normalization is used in this representation to consider the variations in topics. A customized version of the algorithm rather than the standard TF-IDF approach is utilized in BERTopic, allowing for a far better representation. The modified algorithm Grootendorst (2022c) proposed for this computation is shown below.

$$W_{\text{tc}} = \text{tf}_{t,c} \cdot log\left(1 + \frac{A}{\text{tf}_t}\right) \tag{1}$$

Here, the term frequency represents the frequency of term t in class c in Eq. (1). The group of documents combined into a single document for each cluster is class C in this instance. Then, evaluating how much knowledge a term contributes to a class, the inverse document frequency is substituted with the inverse class frequency. It is the frequency of term t across all classes divided by the logarithm of the average number of words per class A. It is beneficial when dealing with extensive collections of documents where it is crucial to identify the most important terms for each class.

In topic modeling, a topic is a cluster of words. These words are selected based on their statistical significance in the model. The importance of words is determined by their frequency of occurrence in the cluster and their co-occurrence with other words in the same context. Therefore, it is very helpful for interpreting the groupings produced by any unsupervised clustering method. Bar charts of the C-TF-IDF scores of each topic provide two insights: Scores for each word in each topic's C-TF-IDF and a comparative study of each topic's distribution. The process used for Topic Modeling is depicted in Fig. 4.

3.2.2 Explainable AI

An emerging subset of AI called explainable AI (XAI) focuses on the readability of ML models. An explainable AI model is a set of steps and strategies that enables one to comprehend and accept ML algorithms' results. It entails specifying the model's correctness and transparency and the outcomes of decision-making assisted by the model. LIME (Ribeiro et al. 2016) is a model-independent interpretability method for individual local predictions. Model agnostic refers to the ability to generate explanations for any DL or ML model. The degree to which a person can decipher the reasoning behind a black box model's choice is known as interpretability. Instead of creating an explanation for the whole, LIME's general operating premise is to localize the problem and explain it.

3.3 Evaluation metrics

The evaluation metrics for classification models are accuracy, precision, recall, and F-score. Accuracy is the ratio of the total number of positives to the total number of classes. Precision is the ratio of true positives to the total number of predicted positives. The recall is the ratio of true positives to the actual number of positives. We can get a harmonic mean of these measures using the F1-Score that considers precision and recall. When there is a significant class imbalance, this is very helpful. The formulas are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2}$$

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

where TN, TP, FN, and FP represent true negative, true positive, false negative, and false positive, respectively in Eq. (2–5).

In general, classification models are assessed using the above metrics, but since data that have been collected and annotated may have bias compared to the real world. Additionally, LIME is employed to explain by enhancing interpretability. In our case, the SWMH data set is based on the mental health-related subreddits where nuances of language and the usage of the words need a deep understanding of the context in which they were used. Therefore, the traditional metrics might not accurately reflect the primary goal of developing the text classification model. In addition to such metrics, analyzing individual predictions using LIME can be helpful. However, there is no formal agreement on what interpretability signifies in ML and its measurement methods (Molnar n.d.). A local prediction is defined as $L(f, g, x)$ applying the regularization parameters $\Omega(g)$. The L indicates the minimized square loss function and g represents the model applied to class G (class G is defined as the set of models that have the capacity for interpretability). The faithfulness of the explanation g to the original Explanation(x) is measured to get an explanation of a local point x. The formula is as follows:

$$\text{Explanation}(x) = \arg\min_{g \in G} L(f, g, x) + \Omega(g) \tag{6}$$

where $\arg\min_g$ is defined as the value of argument *'g'* for which function $GL(f, g, x)$ attains its minimum.

3.4 Model definitions

Multiclass classification, often known as multinomial classification, categorizes cases into three or more classes in statistical classification and ML. Each data sample can be put into a specific class. A data sample, however, cannot concurrently be a member of more than one class. In other words, the classes in multiclass problems are mutually exclusive.

3.4.1 Logistic regression (LR)

By default, logistic regression can only be used to solve binary classification issues. As a result, it has to be modified to enable multiclass categorization issues. Studies (Pranckevičius and Marcinkevičius 2017) have shown that Logistic regression can be used to achieve higher performance in multiclass text classification. Logistic

Regression is helpful in this situation as it uses a sigmoid function to output a probability between zero and one. In this study, we have adapted a logistic regression model to recognize and forecast a multinomial probability distribution known as multinomial logistic regression.

3.4.2 Random forest (RF)

As an extension of bagging, the random forest (Breiman 2001) algorithm randomly chooses portions of the features utilized in each data instance. Decision trees, which are the foundation of the random forest model, are susceptible to class imbalance. Every tree is supported by a "bag," representing a uniform random data sampling. As a result, a class imbalance will, on average, bias each tree in the same direction and magnitude.

3.4.3 Stochastic gradient descent (SGD)

SGD is an optimization approach for updating a model's weights while being trained. The algorithm updates the weights based on the difference between each input text's actual and projected class. By merging various binary classifiers in an "OVA" (one versus all) framework, it can be used for multiclass classification. A binary classifier is learned for each class that can distinguish it from all other classes. For large datasets, SGD provides a quick and effective optimization approach.

3.4.4 Naive Bayes (NB)

Based on the Bayes theorem, the likelihood of a class given a set of features is proportional to the likelihood of the features given the class. Naive Bayes models are frequently employed for text classification problems since they are quick and simple to implement.

3.4.5 XGBoost (XGB)

GradientBoost was first introduced by Chen in 2016 and XGB (Chen and Guestrin 2016) is its improved version. A gradient boosting framework is used by the decision tree-based ensemble ML algorithm XGB to handle missing values in the training phase and control overfitting and split finding.

3.4.6 Long short-term memory (LSTM)

LSTM is a form of recurrent neural network (RNN). For text data sequences where the word order matters, LSTM models are very well suited. Long-term dependencies can be captured by LSTMs, which also have the ability to handle sequential dependencies in the data.

3.4.7 Convolutional neural network (CNN)

Convolutional Neural Networks can be utilized to solve text classification issues in addition to picture and video analysis. When classifying text, the input is handled like a picture, with each word acting as a feature. The fully connected layers are utilized for prediction, whereas the convolutional layers extract features from the input.

3.4.8 BERT

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018) is a pre-trained DL model. BERT models are customized for certain classification tasks after being pre-trained on a sizable corpus of text data. BERT models are ideally suited for text categorization issues since they have a reputation for capturing the context of the words in a text.

3.4.9 MentalBERT

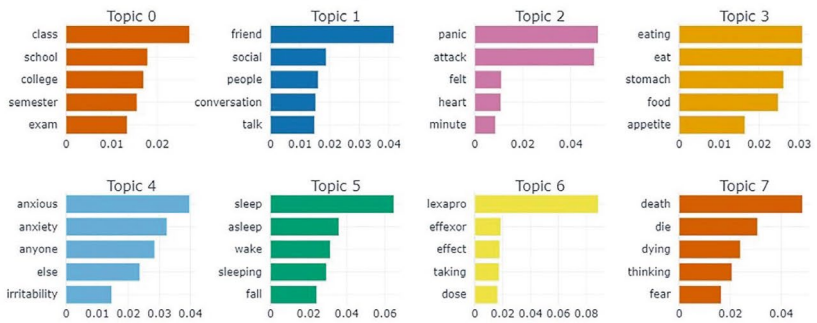
MentalBERT (Ji et al. 2021) uses the BERT model's pre-trained knowledge for contextualized language representations related to mental health. The classification head is trained using a sizable corpus of annotated text to determine the class of a given text document. The MentalBERT model's weights are adjusted during fine-tuning for optimum text categorization performance.

4 Results and discussion

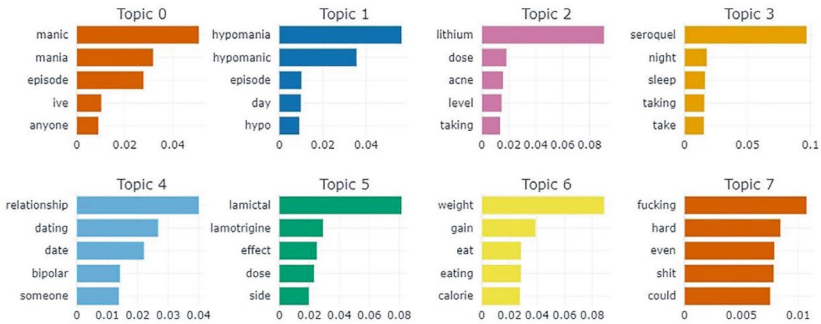
4.1 Topic modeling

We investigated four commonly identified mental health problems, examining the recurring themes and vocabulary that emerged. Plots for the eight most popular topics under each corresponding mental health condition are shown in columns of Fig. 5. Each topic cluster's five most common words are plotted on each bar chart. Each row of the bar graph with the C-TF-IDF score shows the most common terms from each cluster. It can be concluded from the results that the subjects like *class, school, holiday or exam* can trigger the symptoms of mental health problems. These results can be used to aid the therapy and medication processes and to empathize with mental health patients.

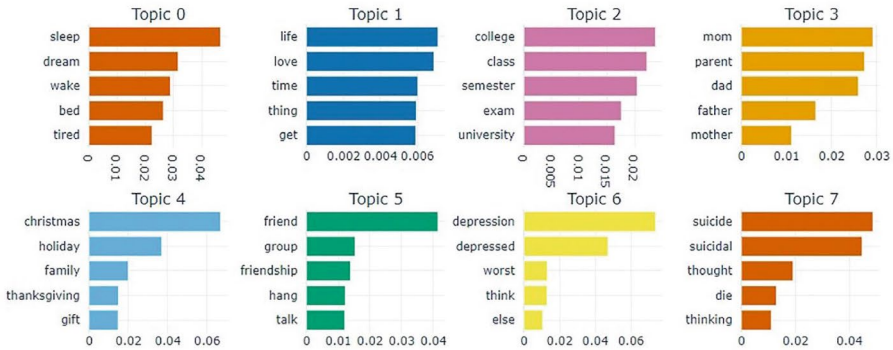
Recurring themes are observed in each class. For instance, topics related to college, work, and relationships frequently appear in every category examined. Demographically speaking, the reason for this can be that Reddit users are mainly young, tech-savvy people who use social media platforms more often to discuss their issues daily. Thus, this study can be extended to focus exclusively on the problems



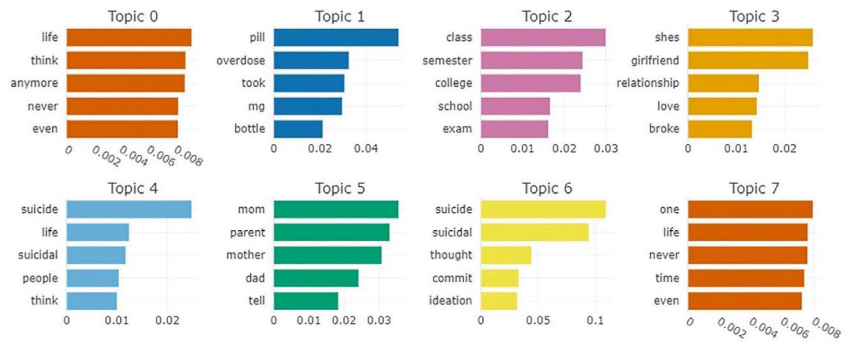
(a) Anxiety



(b) Bipolar



(c) Depression



(d) Suicide

Fig. 5 Graph of recurring words with their C-TF-IDF scores for each class

of youth in modern society that trigger mental health issues and trauma. The themes about bipolar disorder stand out from the rest, as people with bipolar disorder seek professional help compared to the rest. Thus, the topics related to medication and therapy were prominent in that category.

However, it is noteworthy that the BERTopic model has several shortcomings (Grootendorst 2022c). It presumes that each document contains one topic, while in reality, documents may have numerous topics. The word corpus was filtered using the label mental health issue, and five similar motifs in topics were found through qualitative analysis. In each health problem, the topic clusters were related. We could gather subjects that were self-contained and coherent as a result. Given the highly topical nature of Reddit's subreddits, it is expected that some topics, like bipolar disorder studies, are more specific to a particular subreddit than others. The present research facilitates comparing and examining overlapping vocabulary and underlying semantic attributes over time. The relative frequency variation of various clusters can be observed in a given mental health condition, such as depression or suicide. The procedure outlined also enables us to assess conversational shifts related to various topics within a related category and compare it with the rest. It would be helpful for mental health practitioners to understand how various conversations evolve or how they respond to particular events differently.

4.2 Classification

Classification techniques were used to perform Multiclass classification of the text corpus into each category. Since the data set is considered mutually exclusive, each text is only in one category. However, the cases may fall under two categories simultaneously. For instance, a depressed patient might also be suffering from suicidal tendencies at the same time. In addition, in some instances, it was noticed that texts must be of a substantial length to gain enough context to classify correctly.

Using several ML, DL and transformer models for classification allowed us to examine context-based against non-context-based classifier performance. The existence of an imbalanced class distribution within this dataset has repercussions on the model's effectiveness. Because the model exhibits a bias toward the majority class, its capacity to effectively learn from the minority class is compromised. To tackle this challenge, both undersampling and class weight techniques were applied to the dataset. Figure 6 showcases a comparison of the outcomes before and after addressing the data imbalance, providing a visual representation of the effects of these strategies. The domain-specific MentalBERT outperformed all other models (context and non-context-based methods). Table 2 displays the average performance evaluation of multiclass classifiers. The models performed better after undersampling the majority-class data, with the best-performing model MentalBERT achieving 76.70% mean accuracy (best-performing results are in bold in Table 2), outperforming the results published in previous studies (Ji et al. 2022, 2021) by more than 4%. To evaluate the model's classification performance, an accuracy error bar graph (Fig. 7) has been plotted for all nine classification models. Classical Models such as NB, RF, and transformer-based models BERT and MentalBERT have shown very low variance in accuracies compared to CNN and LSTM. This low variance shows that the model's performance is consistent and generalizes well across all the subsets of the dataset. Confusion matrix has also reported in Fig. 8 to understand the statistics of performance for best-performing model. MentalBert has a identify the true labels and false positives and negatives are very less for most of the classes. In literature (Ji et al. 2022), a relation network (RN) based on

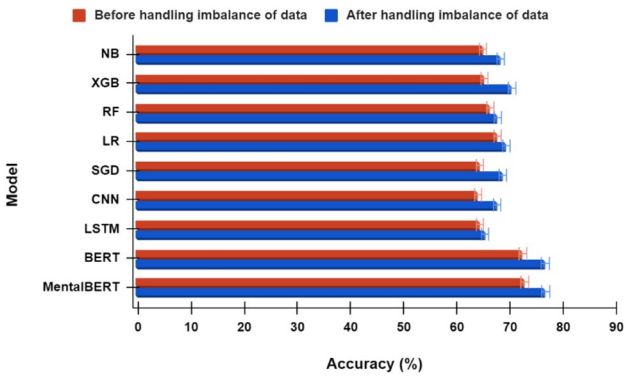


Fig. 6 Comparison of models on the basis of accuracy before and after handling of imbalanced data

Table 2 Results of classification performance metrics for all nine models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
NB	68.23	68.87	68.23	68.32
XGB	70.38	74.49	68.13	70.53
RF	67.69	67.99	67.99	67.40
LR	69.30	69.92	69.30	69.55
SGD	68.63	71.54	68.63	68.77
CNN	67.56	68.54	66.35	67.41
LSTM	65.29	67.40	61.40	64.20
BERT	76.62	76.52	76.56	76.56
MentalBERT	**76.70**	**76.66**	**76.69**	**76.63**

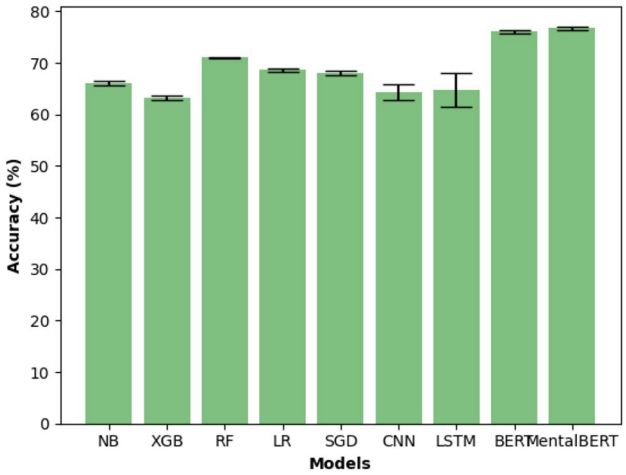


Fig. 7 Error bar graph for accuracy on all nine classification models

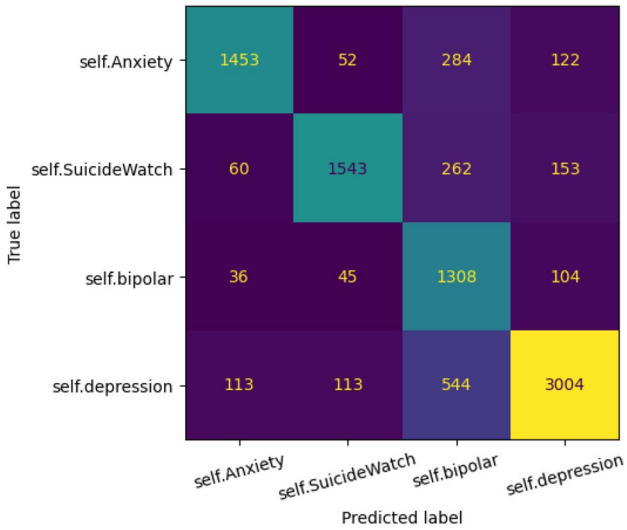


Fig. 8 Confusion matrix for best-performing model (MentalBERT)

the attention technique was used to classify the text. This RN model achieved 64.74% accuracy and 64.78% F1-score. Another study (Ji et al. 2021) applied MentalRoBERTa on the SWMH dataset and attained 72.16% F1-score.

One more experiment was designed to understand further the decision various models took. In this study, nine models are considered for classification. The SGD model outperformed four models (RF, LSTM, XGB and NB) but lagged behind the other four models (BERT, CNN, LR and MentalBERT) in terms of accuracy. Hence, 50 misclassified instances of SGD from all classes were selected and checked for their LIME explanations ("6."). Also, the other 8 models were applied to check their efficiency on these misclassified instances. As shown in Fig. 9, MentalBERT correctly classified twenty-four instances out of 50, whereas RF could correctly classify only sixteen. These posts have mixed topic themes and class-representative keywords that may confuse models, yet MentalBERT performed well compared to all other classification techniques used in this study. Despite having a smaller training corpus size compared to BERT, MentalBERT demonstrates comparable or superior performance on mental health datasets. This is due to MentalBERT's capability for capturing the specific context and sentiments prevalent in mental health-related text. While BERT demands substantial computational resources for training and inference due to its general-purpose applicability and extensive pre-training corpus, MentalBERT's specialized training results in reduced resource utilization during both the training and inference stages. This decreased resource overhead can prove advantageous in situations where computational resources are limited or when deploying the model on devices with constrained memory capacities.

An additional dataset ("depression" 2021) extracted from Facebook comments and posts on mental health-related issues, is analyzed within the same experimental setup to validate the proposed approach. This dataset comprises 6982 social media posts and is publicly available on the Kaggle platform. The dataset consists of two classes and has been annotated using a majority voting approach involving undergraduate students. All nine classification models were applied to this dataset, and the results are presented in Fig. 10. The classification outcomes were compared with the results reported in the literature (Hassan et al. 2021) and Kaggle notebooks to ensure consistency and reliability.

4.3 Explainable AI

LIME determines whether a system produces insightful justifications for its classification. The features employed in this work can be used to support ML assessments and contribute to the design of a reliable user interface. Figure 11 shows the

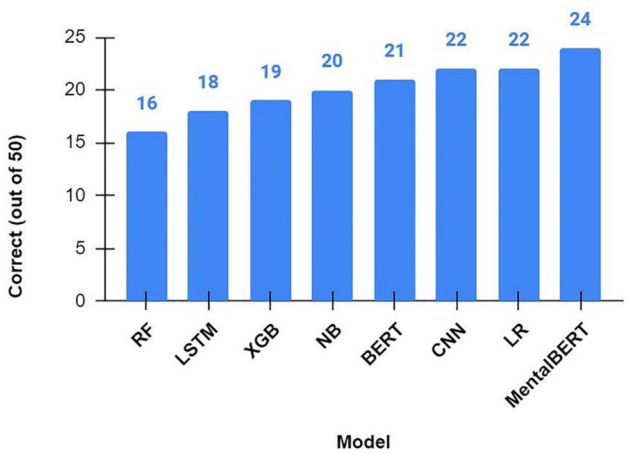


Fig. 9 Count of correct posts classified by all models except SGD

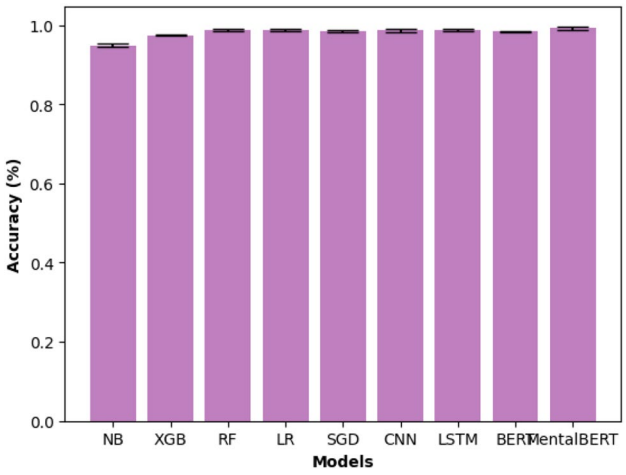


Fig. 10 Error bar graph for accuracy on all nine classification models on the Kaggle dataset

LIME algorithm outcomes for each model. The total class probabilities, the top features with their probabilities, and automatically highlighted features in the sample input text using LIME are considered to understand the results.

The class probabilities for a given model range from 0–0.99%. For instance, as shown in Fig. 11, a test text (I want to turn into an alcoholic and die in 6 years) taken for the depression class for the XGB model gives a 0.52% probability of '*self.suicidewatch*' label because it has the word *die* and the top features with their probabilities explain each class probability indicating the content is biased toward *suicidewatch* class. According to the classification from the XGB and NB models and explanations from LIME, the text consists of *suicidewatch* class (Fig. 11a, b). However, the true class (*self.depression*) of the text sample was not the same as the model's prediction because the sample text has few words (less information) and does not contain class-representative keywords (depression or depressed). The DL—LSTM model performed well for this sample text, while CNN could not because it gets confused with the words like '*deleted'* and '*die'*. RF, LR and SGD also performed well in this text but failed in others (see "6.").

Context-based models have tried to understand the context and give a higher probability to the true class (Fig. 11i). To our knowledge, the class-representative words or repetitive words (Fig. 5) have contributed to the classification of non-context-based models. While context-based models such as BERT and MentalBERT have classified text based on the overall context of the post, not on a few representative words.

The approach used in this study demonstrates a thorough and thoughtful process to enhance the performance and interpretability of NLP models. Firstly, data pre-processing techniques were implemented to clean the input data and reduce noise by removing stopwords, URLs, and punctuation. Additionally, duplicate texts were eliminated to ensure that the model was not biased by redundant information. Data imbalance, which could have led to biased predictions, was addressed through data undersampling and class weight assignment. Undersampling techniques helped balance the distribution of different classes, while class weights ensured that the model received adequate exposure to the minority class. It significantly enhanced the model's capacity to accurately predict all classes, regardless of their prevalence, compared to the previous study (Ji et al. 2021). This approach was further enriched by incorporating BERT for topic modeling. BERT, as a context-aware model, extracted meaningful themes and intricate details from text data related to mental disorders. By harnessing the capabilities of these models, we gained valuable insights into the underlying patterns within the data, thereby advancing our comprehension of mental disorders and their subtleties. Employing LIME for an explainable NLP analysis bolstered the transparency and reliability of this approach. LIME enabled us to analyze and elucidate the model's predictions by pinpointing crucial features and elucidating their contributions. It facilitated a deeper understanding of the model's decision-making process, rendering it more understandable and trustworthy in contrast to the prior study (Ji et al. 2021). Our approach amalgamates data pre-processing, the rectification of class imbalance, the utilization of BERT for topic modeling, the application of LIME for explainable NLP, culminating in a comprehensive workflow that not only enhances the model's predictive performance but also furnishes transparency in its results.

5 Conclusion and future work

This work captured complex relationships in a wide range of textual data, notably social media posts, using NLP techniques. It examined the distribution of topics in the Reddit-based online community for mental health. The social media data was analyzed for the significant themes related to mental health issues using BERTopic. The topics like relationships, exams, and school were found to impact mental health conditions directly. In addition, the themes that emerged from each category's topics were found to fit into several identifiable patterns. A few patterns related to school, friends and exams are recurring, frequently employing vocabulary from dialogues about mental health. The differences and similarities among the themes covered by the corpus of texts in each community were examined.

A comparative analysis using nine classical state-of-the-art classification techniques was done to classify the mental health disease that may help professional mental health therapists by automating the textual analysis

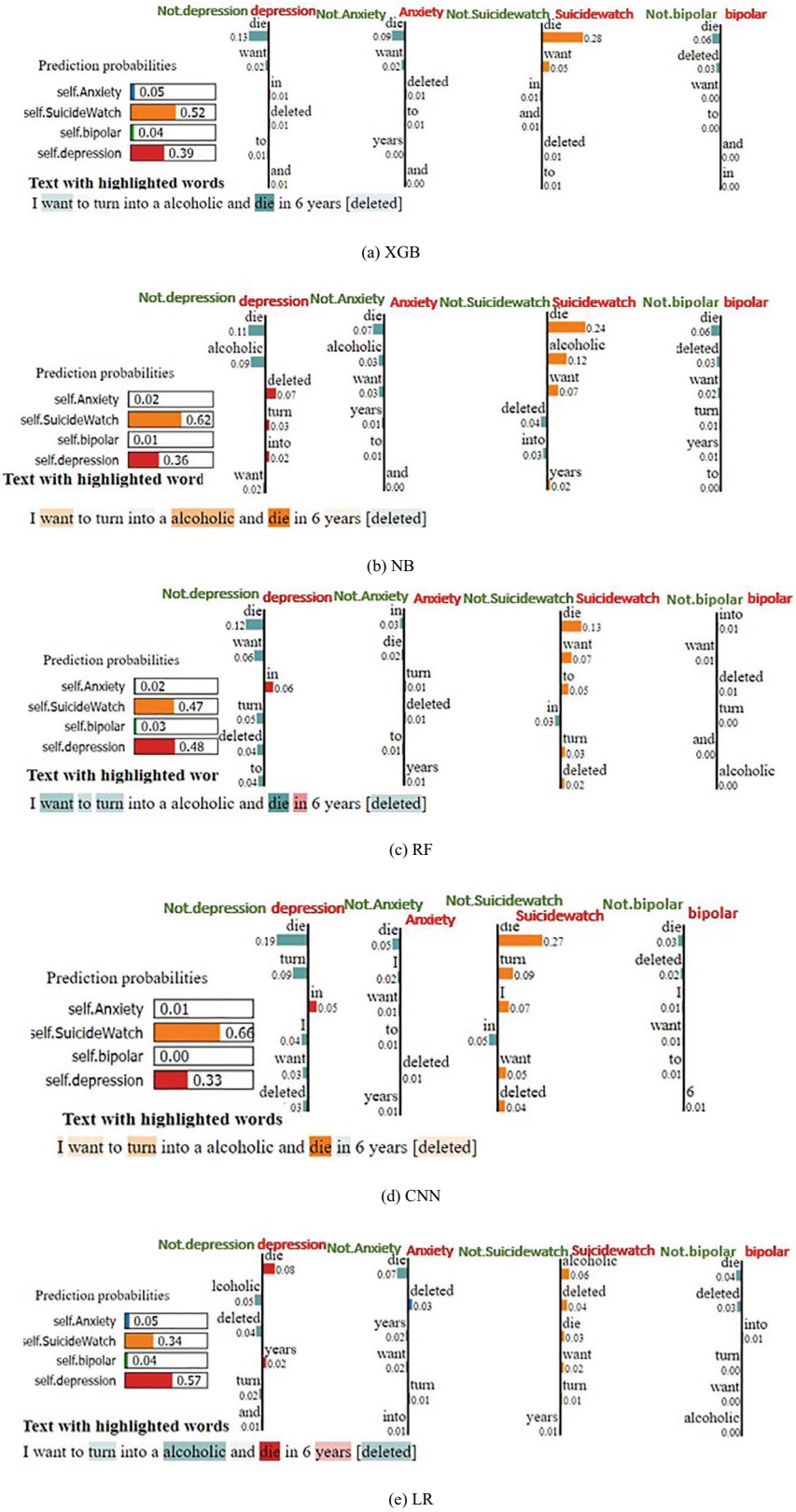
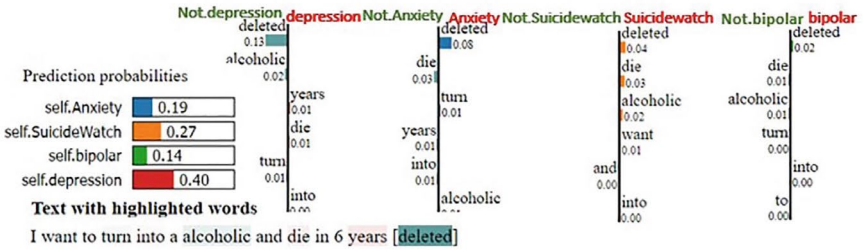
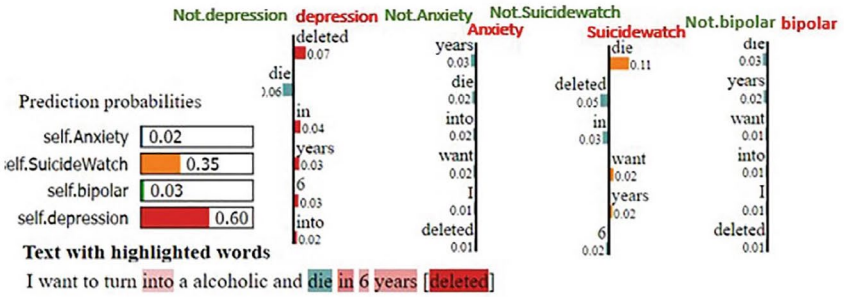


Fig. 11 Prediction probabilities of all nine models using LIME technique to explain model's decision

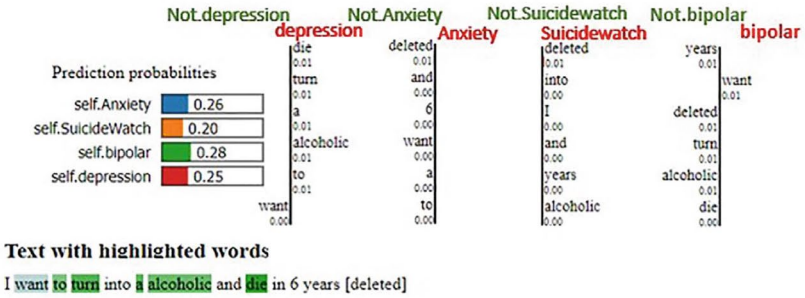
Fig. 11 (continued)



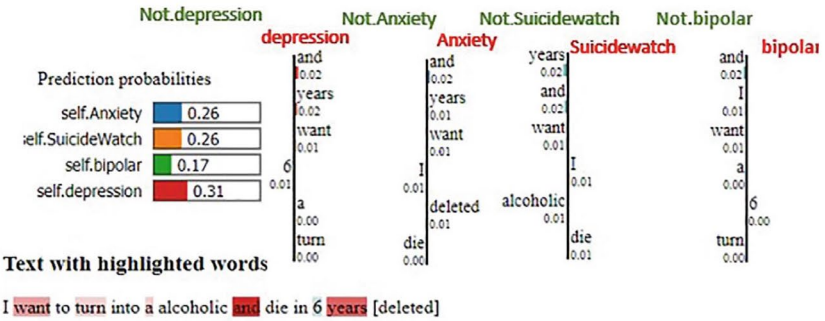
(f) SGD



(g) LSTM



(h) BERT



(i) MentalBERT

process. The transformer model, MentalBERT performed the best, considering the significance of context, and achieved 76.70% accuracy, significantly higher than the reported accuracies. The performance of the models and techniques was evaluated using the explainable AI technique–LIME. It aids in clarifying why a model is producing particular predictions and fostering trust. The results from this study can help identify users who may be at threat of mental health disease. This approach can differentiate users with potential mental health diseases, namely bipolar, depression, anxiety, and even suicidal ideation, based on symptoms extracted from their posts. The recognition of these recurring patterns and themes holds significant potential to assist mental health professionals in their practice. By identifying language patterns that signal the onset of mental health challenges, professionals can intervene proactively and extend support to individuals facing such risks. This early intervention has the capacity to curtail the progression of these mental challenges into more severe conditions. Mental health professionals can apply these insights to fine-tune their interventions and outreach efforts, increasing their effectiveness. Moreover, this study can serve as a valuable resource for policymakers, furnishing crucial data to shape mental health policies and initiatives. Comprehending the central themes and issues impacting mental health, as elucidated in this study, can inform the development of targeted mental health programs. In this research, we utilized solely English-language Reddit posts as our primary social media data source, which represents a constraint in our study. Enhancing the findings could involve incorporating information from a broader array of social media platforms beyond those using English, as well as interactions with individuals from a more extensive range of socio-economic backgrounds. This approach can be applied to larger datasets and more variable investigations in the future to improve the generalization of the study. It can be developed to explore the relationships between the uncovered topics related to mental health, like potential suicide identification or drug addiction.

Appendix 1

A sample text (lets called 'A') '*Feelings of excitement when thinking about suicide The past few days I've been dwelling on stupid past mistakes and getting a lot of invasive suicidal thoughts, but lately when I get them I start to feel excited like I'm going some place like an amusement park or something. When I think about killing myself it feels so right, rather than feeling dark and wrong. My heart starts racing really fast and I get tunnel vision at times and think about hanging myself, but then I immediately calm down and the thoughts stop after just 10 s or so. I'm not sure if this is common or not.*' is taken from test set belongs to *self.depression* class has classified for all nine classification models. In Table 3, classification insights have been reported for all the models used in this study. XGB, NB, CNN, SGD and LSTM wrongly classified the sample text as self.SuicideWatch (because it contains keyword *suicide*), self.Anxity(it contains words *racing* and *calm*), self.bipolar, self.SuicideWatch, and self.Anxity, respectively. However, RF, LR, BERT, MentalBERT classified the this text correctly as *self.depression*.

The true class for another sample text (lets called 'B') '*Back Well, after my close call that was the subject of my first post, here, I vowed to do something with my life, but it looks like I've gone around in a big fucking circle. I've liked her since high school, I've spent years becoming best friends with her and her boyfriend and she just doesn't fuckin want it. There is literally no end to this Romeo style Petrarchan lover bullshit but as much as I'm self-aware and introspective, I'm also a slave to my biochemical pitfalls. I'm seriously considering it.'* is self.SuicideWatch. Non-context-based models XGB and LR correctly identified this text where as deep learning models CNN and LSTM both misclassified sample text 'B.' Both context-based models BERT and MentalBERT also predicted actual class as shown in Table 4. Although BERT and MentalBERT gave same prediction probabilities for all classes but the selected different words to predict true class.

One more sample text (lets called 'C') from class self.depression has been taken '*depression and anxiety Attack!! Earlier I was feeling depressed and now I am having a anxiety attack. my stomach is churning, and I just want to lay down and die in all Honesty I feel like bashing my head against a wall maybe then I wouldn't feel like this. I have a sense of dread. I don't have any clue why I am feeling this way, but I just want to cry myself to sleep at least. but no. that's not even a possibility because I can't stop fidgeting and stirring and my mind won't calm down. not even breathing in and out helps.'* to understand the model's classification decision. The text contains both anxiety and

Table 3 Lime explanation for sample text A of each model

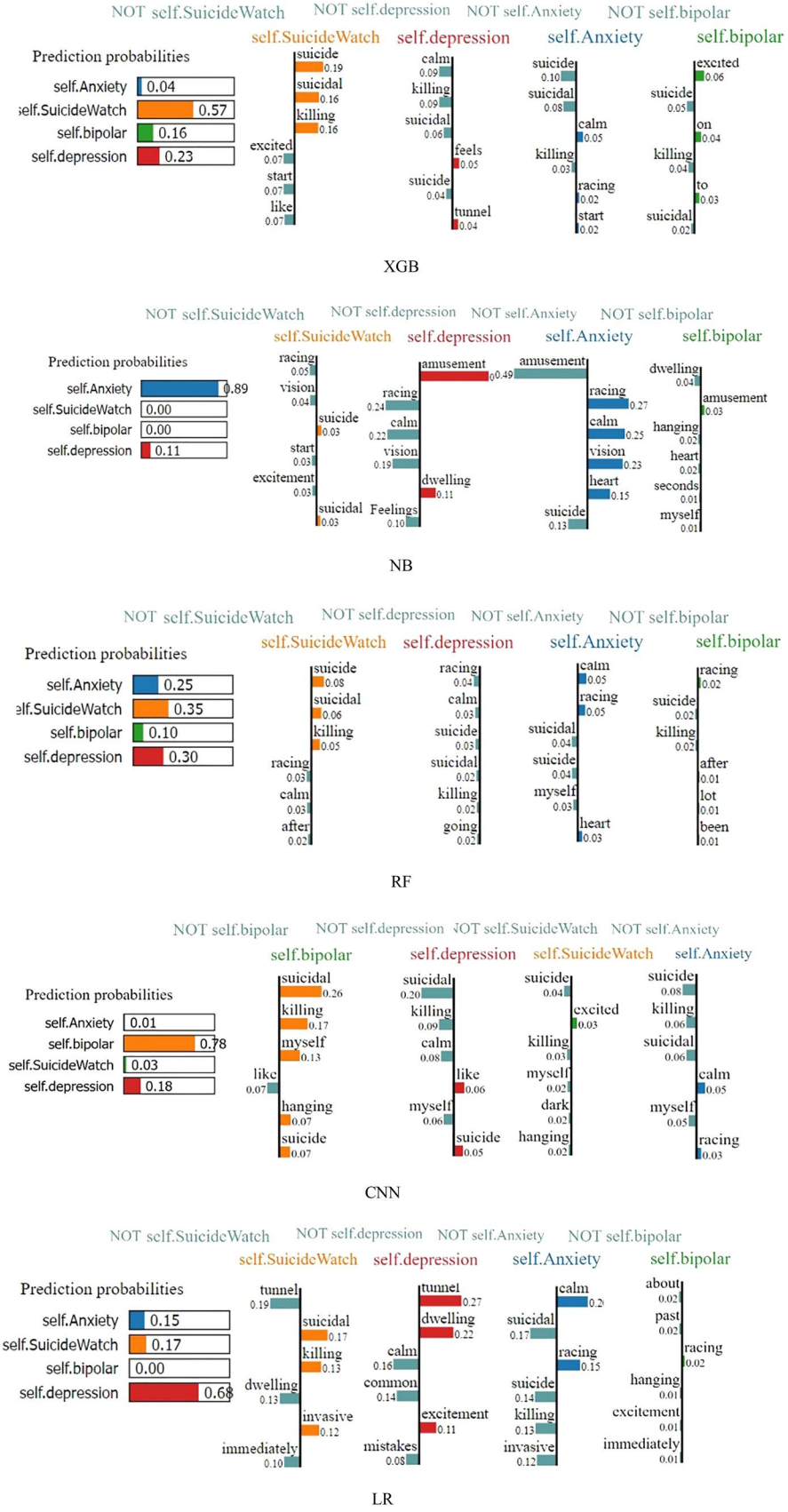


Table 3 (continued)

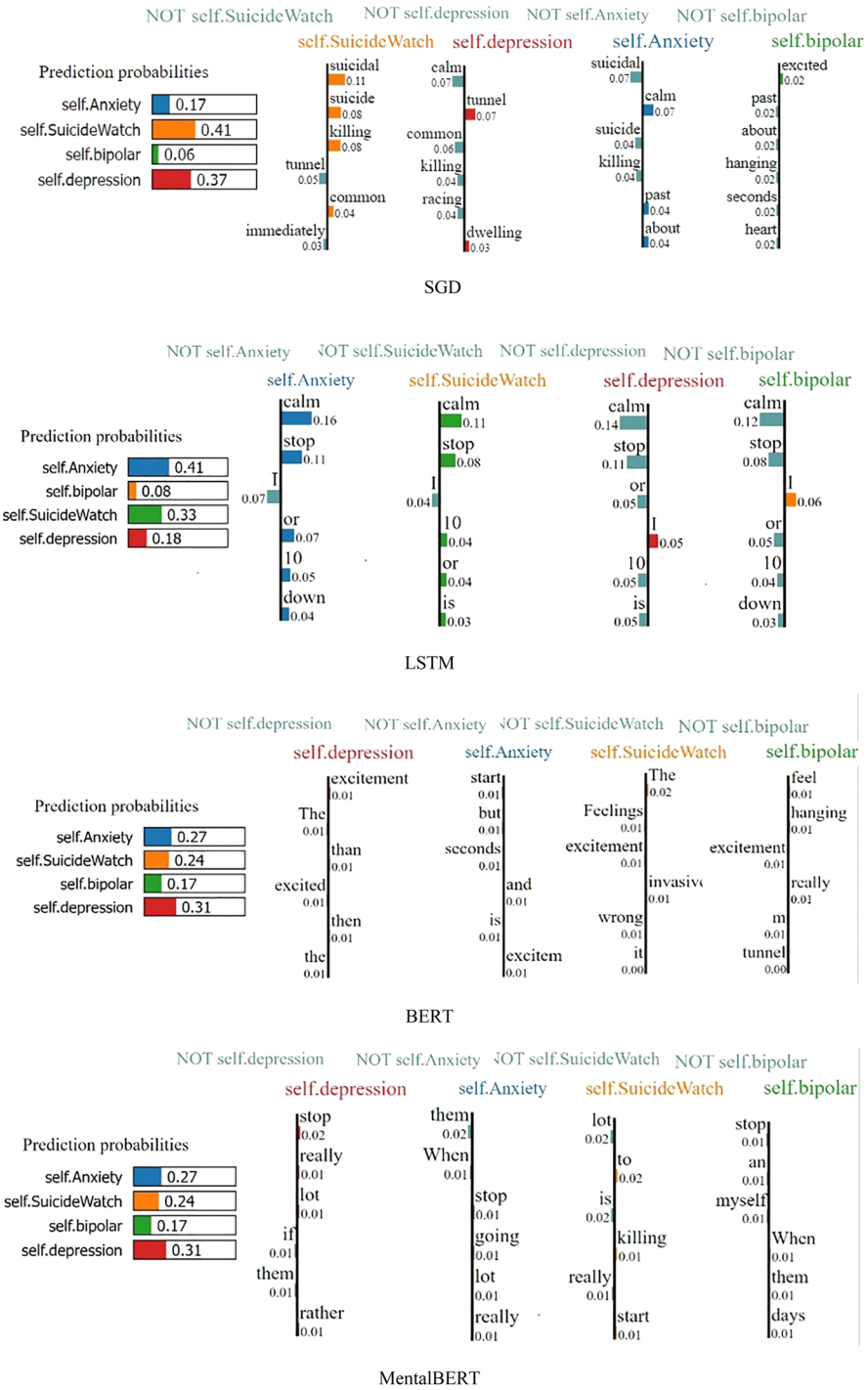


Table 4 Lime explanation for sample text B of each model

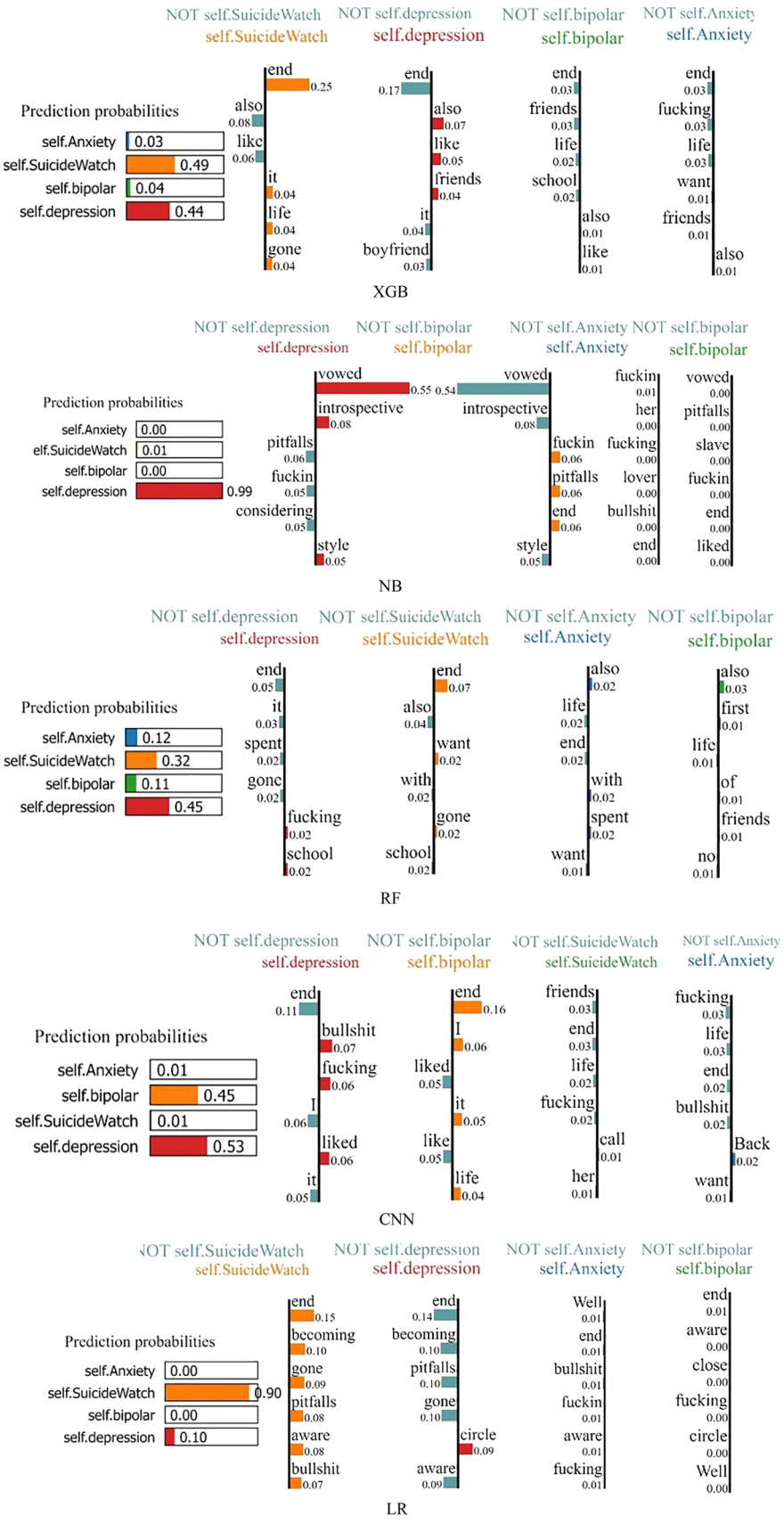


Table 4 (continued)

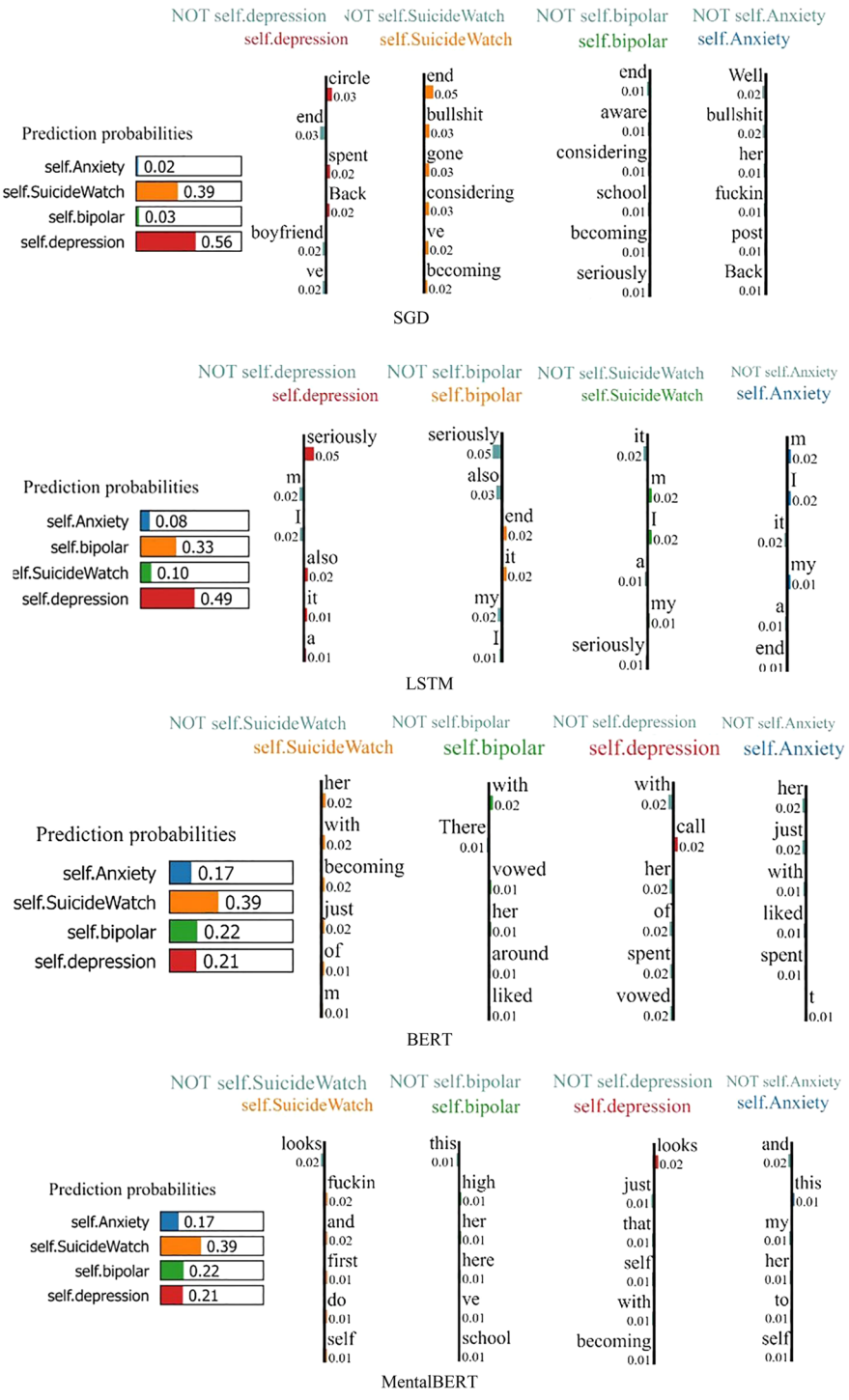


Table 5 Lime explanation for sample text C of each model

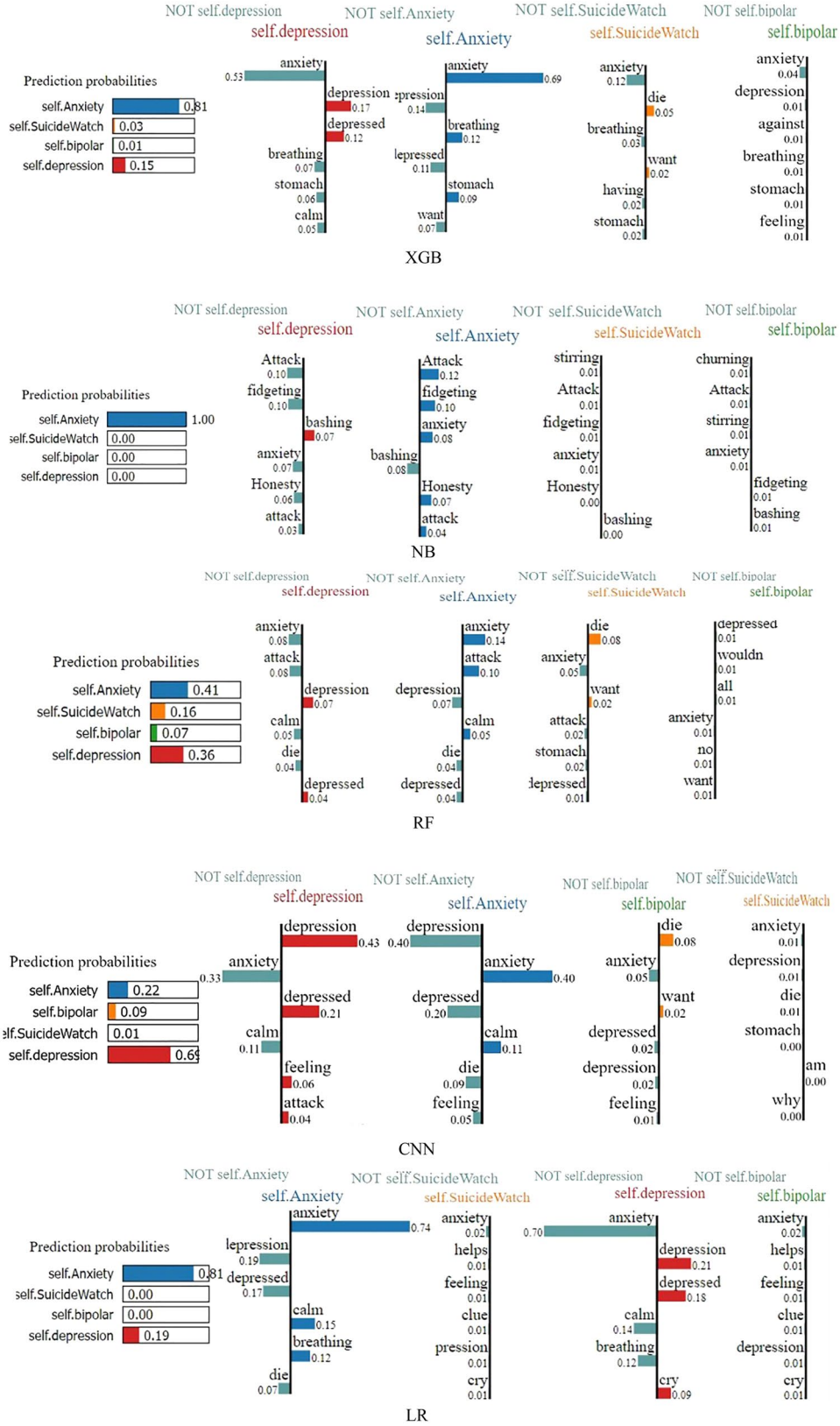
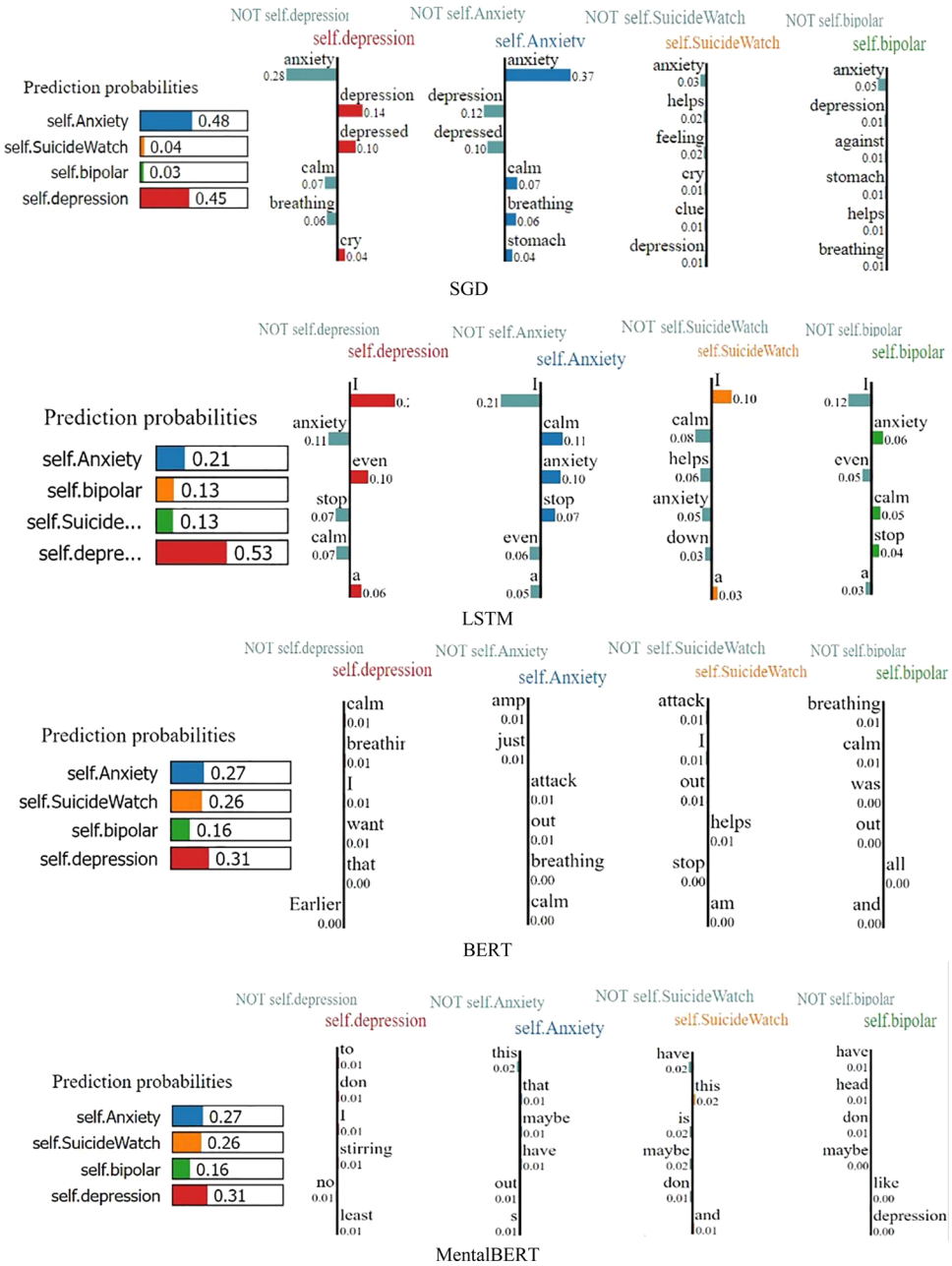


Table 5 (continued)



depression related words. Therefore, models XGB, NB, RF, SGD and LR predicted class self.Anxiety (Table 5). However, deep learning and transformer models CNN, LSTM, BERT AND MentalBERT classified actual class.

Acknowledgements The authors would like to thank Project SAMARTH, an initiative of the Ministry of Education (MoE), Government of India, at the University of Delhi South Campus (UDSC), for their support.

Authors contributions PDT was contributed to methodology and writing—original draft, NA was contributed to conceptualization, methodology, visualization, validation, and writing—original draft, VV was contributed to methodology and writing—original draft, GJS was contributed to conceptualization, visualization, validation, and writing—review and editing, **SS** was contributed to writing—review and editing, AP was contributed to validation and writing—review and editing.

Funding No funding was received for conducting this study.

Data availability The dataset analyzed in the present study can be available on request using the link: https://doi.org/10.5281/zenodo.6476179.

Code availability The two models of BERT and MentalBERT for this dataset have been released in Hugging Face hub and they can be downloaded for further studies.

MentalBERT: https://huggingface.co/tiya1012/swmh4_mtb
BERT: https://huggingface.co/tiya1012/swmh4_bert

Declarations

Conflict of interests The authors have declared that they have no conflict of interest.

Ethics approval and consent to participate Not applicable.

References

Abuzayed A, Al-Khalifa H (2021) BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique. Procedia Computer Science 189:191–194. https://doi.org/10.1016/j.procs.2021.05.096

Alotaibi W, Alomary F, Mokni R (2023) COVID-19 vaccine rejection causes based on Twitter people's opinions analysis using deep learning. Soc Netw Anal Min 13:62. https://doi.org/10.1007/s13278-023-01059-y

Benrouba F, Boudour R (2023) Emotional sentiment analysis of social media content for mental health safety. Soc Netw Anal Min 13:17. https://doi.org/10.1007/s13278-022-01000-9

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of machine Learning research, 993–1022.

Boettcher N (2021) Studies of Depression and Anxiety Using Reddit as a Data Source: Scoping Review. JMIR Ment Health 8(11):e29487. https://doi.org/10.2196/29487

Breiman L (2001) Random Forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. ArXiv DOI 10(1145/2939672):2939785

Dao B, Nguyen T, Venkatesh S, Phung D (2015) Nonparametric discovery of online mental health-related communities. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, France, pp 1–10. https://doi.org/10.1109/DSAA.2015.7344841

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint https://doi.org/10.48550/arXiv.1810.04805

"depression", Kaggle.com, 2021, [online] Available: https://www.kaggle.com/datasets/sahasourav17/students-anxiety-and-depression-dataset.

Garg M, Saxena C, Krishnan V, Joshi R, Saha S, Mago V, Dorr BJ (2022) CAMS: an annotated corpus for causal analysis of mental health issues in social media posts. arXiv preprint. https://doi.org/10.48550/arXiv.2207.04674

Gemmell J, Isenegger K, Dong Y, Glaser E, Morain A (2019) Comparing Automatically Extracted Topics from Online Mental Health Disorder Forums. In:International Conference on Computational Science and Computational Intelligence (CSCI), pp 1347–1352. https://doi.org/10.1109/CSCI49370.2019.00252

Gkotsis G, Oellrich A, Velupillai S, Liakata M, Hubbard TJ, Dobson RJ, Dutta R (2017) Characterisation of mental health conditions in social media using Informed Deep Learning. Sci Rep 7:45141. https://doi.org/10.1038/srep45141

Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint. https://doi.org/10.48550/arXiv.2203.05794

Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ (2019) XAI—Explainable artificial intelligence. Sci Robot,4(37). https://doi.org/10.1126/scirobotics.aay7120

Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC (2017) Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci 18:43–49. https://doi.org/10.1016/j.cobeha.2017.07.005

Hanna F, Barbui C, Dua T, Lora A, van Regteren AM, Saxena S (2018) Global mental health: how are we doing? World Psychiatry 17(3):367–368. https://doi.org/10.1002/wps.20572

Hassan MM, Khan MAR, Islam KK, Hassan MM, Rabbi MMF (2021) Depression Detection system with Statistical Analysis and Data Mining Approaches. In: International Conference on Science & Contemporary Technologies (ICSCT), Dhaka, Bangladesh, pp 1–6, https://doi.org/10.1109/ICSCT53883.2021.9642550.

Hu Y, Sokolova M (2021) Explainable multi-class classification of the camh covid-19 mental health data. arXiv preprint. https://doi.org/10.48550/arXiv.2105.13430

Huang X, Wang S, Zhang M, Hu T, Hohl A, She B, Gong X, Li J, Liu X, Gruebner O, Liu R, Li X, Liu Z, Ye X, Li Z (2022) Social media mining under the COVID-19 context: Progress, challenges, and opportunities. International Journal of Applied Earth Observation and Geoinformation: ITC Journal 113:102967. https://doi.org/10.1016/j.jag.2022.102967

Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A (2018) Depression detection from social network data using machine learning techniques. Health Inf Sci Syst 6(1):8. https://doi.org/10.1007/s13755-018-0046-0

Ji S, Li X, Huang Z, Cambria E (2022) Suicidal ideation and mental disorder detection with attentive relation networks. Neural Comput Appl 34(13):10309–10319. https://doi.org/10.1007/s00521-021-06208-y

Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E (2021) MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. arXiv https://doi.org/10.48550/arXiv.2110.15621

Kamarudin NS, Beigi G, Liu H (2021) A study on Mental Health Discussion through Reddit. In: International conference on software engineering and computer systems and 4th international conference on computational science and information management, ICSECS-ICOCSIM. https://doi.org/10.1109/ICSECS52883.2021.00122

Kathy L, Agrawal A, Choudhary A (2015) Mining Social Media Streams to Improve Public Health Allergy Surveillance. In:IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'15), pp-815–822. https://doi.org/10.1145/2808797.2808896

Kilbourne AM, Beck K, Spaeth-Rublee B, Ramanuj P, O'Brien RW, Tomoyasu N, Pincus HA (2018) Measuring and improving the quality of mental health cara global perspective. World Psychiatry 17(1):30–38. https://doi.org/10.1002/wps.20482

Kim J, Lee J, Park E, Han J (2020) A deep learning model for detecting mental illness from user content on social media. Sci Rep 10:11846. https://doi.org/10.1038/s41598-020-68764-y

Kotenko I, Sharma Y, Branitskiy A (2021) Predicting the Mental State of the Social Network Users based on the Latent Dirichlet Allocation and fastText. In:IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), pp 191–195. https://doi.org/10.1109/IDAACS53288.2021.9661061

Lin YS, Tai LK, Chen AL (2023) The detection of mental health conditions by incorporating external knowledge. J Intell Inf Syst. https://doi.org/10.1007/s10844-022-00774-w

Liu D, Feng XL, Ahmed F, Shahid M, Guo J (2022) Detecting and measuring depression on social media using a machine learning approach: systematic review. JMIR Ment Health 9(3):e27244. https://doi.org/10.2196/27244

Molnar, C (2022) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2. https://christophm.github.io/interpretable-ml-book

Pranckevi˘cius T, Marcinkevi˘cius V (2017) Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Baltic J Modern Comput 5(2):221

Qi Y, Shabrina Z (2023) Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach. Soc Netw Anal Min 13:31. https://doi.org/10.1007/s13278-023-01030-x

Ren L, Lin H, Xu B, Zhang S, Yang L, Sun S (2021) Depression detection on reddit with an emotion-based attention network: algorithm development and validation. JMIR Med Inf 9(7):e28754. https://doi.org/10.2196/28754

Ribeiro MT, Singh S, Guestrin C (2016) Why Should I Trust You?: Explaining the Predictions of Any Classifier. arXiv preprint. https://doi.org/10.48550/arXiv.1602.04938

Rizvi STR, Ahmed S, Dengel A (2023) ACE 2.0: A Comprehensive tool for automatic extraction, analysis, and digital profiling of the researchers in scientific communities. Soc Netw Anal Min 13:81. https://doi.org/10.1007/s13278-023-01085-w

Saha B, Nguyen T, Phung D, Venkatesh S (2016) A framework for classifying online mental health-related communities with an interest in depression. IEEE J Biomed Health Inf 20(4):1008–1015. https://doi.org/10.1109/JBHI.2016.2543741

Sangaraju VR, Bolla BK, Nayak DK, Kh J (2022) Topic modelling on consumer financial protection bureau data: an approach using BERT based embeddings. arXiv preprint. https://doi.org/10.48550/arXiv.2203.05794

Saxena C, Garg M, Ansari G (2022) Explainable causal analysis of mental health on social media data. Explainable causal analysis of mental health on social media data. arXiv preprint. https://doi.org/10.48550/arXiv.2210.08430

Stein DJ, Palk AC, Kendler KS (2021) What is a mental disorder? An exemplar focused approach. Psychol Med 51(6):894–901. https://doi.org/10.1017/S0033291721001185

Suicide data: Mental Health and Substance Use (2021). https://www.who.int/teams/mental-health-and-substance-use/data-research/suicide-data Accessed 5 January 2023.

Verma R, Chhabra A, Gupta A (2023) A statistical analysis of tweets on covid-19 vaccine hesitancy utilizing opinion mining: an Indian perspective. Soc Netw Anal Min 13:12. https://doi.org/10.1007/s13278-022-01015-2

Wainberg ML, Scorza P, Shultz JM et al (2017) Challenges and opportunities in global mental health: a research-to-practice perspective. Curr Psychiatry Rep 19(5):28. https://doi.org/10.1007/s11920-017-0780-z

World mental health report: Transforming mental health for all - executive summary (2022). https://www.who.int/publications/i/item/9789240049338 Accessed 28 December2022.

Yazdavar AH, Mahdavinejad MS, Bajaj G, Thirunarayan K, Pathak J, Sheth A (2018) Mental health analysis via social media data. In: IEEE international conference on healthcare informatics (ICHI), NY, USA, pp 459–460. https://doi.org/10.1109/ICHI.2018.00102

Zanwar S, Wiechmann D, Qiao Y, Kerz E (2022) Exploring Hybrid and Ensemble Models for Multiclass Prediction of Mental Health Status on Social Media. arXiv preprint. https://doi.org/10.48550/arXiv.2212.09839

Zhou J, Zogan H, Yang S, Jameel S, Xu G, Chen F (2021) Detecting community depression dynamics due to covid-19 pandemic in Australia. IEEE Trans Comput Soc Syst 8(4):982–991. https://doi.org/10.1109/TCSS.2020.3047604

INTERNSHIP REPORT

Submitted by
Harshita Sharma
harshita.sharma@iic.ac.in

Internship Period:
August 2022 to February 2023

Cyber Physical Systems Laboratory,
Institute of Informatics and Communication,
University of Delhi South Campus

Table of contents

1. Acknowledgement

I extend my sincere gratitude to the Cyber-Physical Systems Lab at IIC for providing me with the invaluable opportunity to undertake this six-month internship. Thanks to all the mentors, for their unwavering support, valuable insights, and constant encouragement throughout the internship. I would also like to express my appreciation to the entire team at the CPS Lab for fostering a collaborative and enriching work environment.

2. Tasks

2.1. Task -1 :- Website maintenance

This task was assigned in order to keep the official website for Cyber-Physical Systems Lab at IIC maintained and Updated. This multifaceted responsibility involved not only routine tasks but also to optimize performance. This task involved following contributions on regular basis:-

- Content Management and regularly updating articles to ensure the latest information was accessible to the website's visitors.
- Introducing new pages to the website such as section for lab domain articles and news articles.
- Implementing a systematic approach to content review, ensuring accuracy and relevance of the articles contributed by the lab members.
- Collaborating with subject matter experts to incorporate their insights into the articles.
- Conducting periodic checks to ensure the overall functionality and implementing updates to the website's infrastructure.

Link - https://cps.iic.ac.in/ ;

2.2. Task -2 :- Internship banner design

The task was to design a banner for Internship advertisement that was published on the official website of University of Delhi and at various other platforms.

Link - canva

2.3. Task -3 :- UI for Online University

The task was to conceptualize and design a basic user interface for an Online University. This task involved following contributions on regular basis:-

- Participating in brainstorming sessions to contribute creative ideas that aligned with the educational platform's objectives and user needs.
- Utilizing Figma as the primary design tool to create wireframes and high-fidelity prototypes.

Link - Figma

2.4. Task -4 :- UI improvement suggestion for Employee dashboard Samarth portal

The task was to provide valuable input for enhancing the User Interface of the Samarth Employee Dashboard. This task included the following contribution:-

- Conducted an in-depth analysis and prepared a UI case study,
- Suggested improvements to address usability issues, streamline navigation, and optimize the overall user journey within the dashboard.

Link - Copy of Copy of UI case study:- Samarth Employee Dashboard

2.5. Task -5 :- Front End for Smart Concerned Classes

This task included the design and implementation of the user interface for Smart Concerned Classes. This involved UI design and frontend development for the web application. Continuous discussions with the team regularly regarding progress, addressing challenges, and aligning development efforts with project objectives also took place. Engaging in discussions during meetings to address challenges encountered in the development process.

Link - Github ; Figma

2.6. Task -6 :- DLS Posters

The task was to design promotional posters for the Distinguished Lecture Series (DLS). These posters contributed to the overall success of the DLS events.

Link - Figma

2.7. Task -7 :- Diagrams and Infographics

The task was to contribute to the research process by designing diagrams and infographics for some research papers of the lab. This task includes some key responsibilities such as:-

- Collaborating with researchers and authors to comprehend the concepts and technical details presented in the research papers.
- Ensuring that the visual representations aligned with the tone and goals of the respective research papers.
- Asking for validation of visual representations from the author and incorporating necessary changes.

Link - Folder Link

2.8. Task -8 :- Chatbot Development Task

The task of the development of a Chatbot consisted of the study regarding the integration of chatbots in higher education management systems. The primary objective was to explore how Chatbots could streamline and enhance the admission procedures for students. The task included following subtasks:-

- Conducting a thorough literature review to understand existing implementations of Chatbots in education and higher education management systems.
- Identifying best practices, challenges, and potential areas of improvement in leveraging Chatbots for admission procedures.
- Engaging with educational experts, administrators, and admission personnel to gather insights into the specific needs and pain points in the admission process.
- Simultaneously, developing the Chatbot with a specific focus on addressing the identified needs in the admission procedure.
- Integrating functionalities into the Chatbot that could provide information, guide students through the admission process, and answer frequently asked questions.

By aligning the Chatbot's functionalities with the insights gained from the study, the project sought to make a positive impact on the efficiency of admission procedures in educational institutions.

Link - Github

2.9. Task -9 :- AR Lab Prototype

The task was the initiation of an augmented reality (AR) version for the lab. The primary objective was to craft a prototype of the AR lab, offering augmented reality experience.

To inform the development process, a comprehensive survey was taken to explore the integration of WebXR, AR by Google ARCore, and app development. The objective was to assess the most effective approach for implementing augmented reality in the lab context. After careful consideration, a decision was made and a prototype was made to develop a web application using Three.js and implementing WebXR for an immersive and accessible experience.

Link - Github

3. Summary

My six-month internship at the CPS Lab has been a transformative experience, contributing significantly to my professional development. Through diverse tasks and projects, I gained hands-on experience in various aspects of web development, UI/UX design, and graphic design.

The assigned tasks enabled me to enhance a wide range of skills. From website maintenance and banner design to UI/UX development and chatbot implementation, I acquired practical knowledge and proficiency in relevant tools and technologies. Overcoming the challenges that occurred during the span of the internship expanded my technical skills and deepened my appreciation for user-centric design principles.

Contributing to projects like the Front End for Smart Concerned Classes allowed me to collaborate with diverse teams which improved my communication skills, adaptability, and understanding of collaborative tools. The tasks of maintaining the official website and YouTube channel of the lab and designing banners for internship promotions and Distinguished Lecture Series contributed to the overall success of the digital presence of the lab.

In conclusion, I am deeply grateful for the enriching experiences and supportive environment provided by the organization. This internship has not only equipped me with practical skills but has also shaped my professional identity and aspirations. I look forward to apply these learnings in my future career path.

WADCAT: Web Application for Data Collection and Analysis of Topics

1st Abhishek Shukla
Instititute of Informatics and Communication
University of Delhi South Campus
New Delhi, India
abhishek.shukla@iic.ac.in

***Abstract*—Web application development has become increasingly critical in the context of data collection and analysis, necessitating efficient solutions to handle large-scale data processing. This paper introduces WADCAT (Web Application for Data Collection and Analysis of Topics), a highly scalable web application designed to address the challenges associated with data retrieval and analysis from diverse sources. It is leveraging APIs from prominent platforms such as Google, Reddit, and Twitter, enabling seamless data collection and storage in CSV format. Through advanced preprocessing techniques, including comprehensive natural language processing, the application ensures data cleanliness and suitability for subsequent analysis. By harnessing pre-trained models, it empowers users to perform sophisticated topic modeling and generates comprehensive reports with diverse graph types, facilitating meaningful insights. Equipped with a user-friendly interface, WADCAT simplifies the entire data collection, preprocessing, and analysis workflow, reducing effortless utilization by users. However, recognizing the need for customizability, future work could focus on developing tailored models to accommodate specific dataset requirements. WADCAT emerges as a compelling solution for robust data collection and analysis, poised to make significant contributions to the field and address the pressing demands for efficient web application solutions in the era of big data.**

***Index Terms*—Web Application Development, Data Collection, Data Analysis, Topic Modeling**

I. Introduction

In today's data-driven world, organizations face challenges in efficiently collecting and analyzing data from multiple sources. The complexity of data collection, the need for accurate analysis techniques like Topic Modeling, and the lack of comprehensive reporting capabilities hinder effective decision-making. To address these challenges, we introduce WADCAT (Web Application for Data Collection and Analysis of Topics). WADCAT is designed to provide a scalable, user-friendly platform that streamlines data collection, enables accurate analysis through topic modeling, and generates insightful reports. With WADCAT, organizations can overcome the limitations of existing solutions and make data-driven decisions more efficiently.

A. Solving Data Collection and Analysis Challenges

WADCAT offers a range of features to address data collection and analysis challenges. Firstly, it optimizes the data collection process by utilizing APIs (Application Program Interfaces) from various sources, such as Google, Reddit, and Twitter, to streamline data retrieval. This automated approach significantly reduces the time and effort required for data collection. Secondly, WADCAT eliminates the need for manual coding with its user-friendly graphical interface, empowering users to configure data collection parameters and customize analysis options effortlessly. By removing the coding barrier, WADCAT enables organizations to leverage their capabilities without extensive programming expertise. Thirdly, WADCAT ensures efficient storage and scalability through cloud technologies, allowing organizations to seamlessly scale their data infrastructure as their needs grow.

B. Overview of WADCAT

WADCAT offers several key features and benefits to enhance data collection and analysis. It provides cross-platform compatibility, enabling users to access and utilize the application across different operating systems. The deployment and update process is simplified, allowing organizations to quickly set up and benefit from the latest features and enhancements. With scalability at its core, WADCAT accommodates increasing data volumes while maintaining optimal performance. The application's accessibility ensures that users with varying technical expertise can effectively navigate and utilize its functionalities. By reducing development efforts and minimizing reliance on costly third-party applications, WADCAT presents a cost-effective solution for organizations. WADCAT's architecture promotes easier maintenance and support, ensuring consistent availability and minimizing downtime, incorporating analytics capabilities to provide valuable insights into user behaviour and data trends, and facilitating data-driven decision-making.

II. Literature Review

Data collection and analysis play a vital role in various fields, providing valuable insights and informing decision-making processes. With the proliferation of online platforms and the availability of vast amounts of data, web applications have emerged as essential tools for enhancing data collection and analysis. Researchers have focused on various aspects of web application-based data collection, emphasizing the importance of efficient algorithms and scalable architectures. V. Singrodia et al. (2019) highlighted the significance of web scraping techniques to collect data from websites and

social media platforms. They emphasized the need for proper organization and management of collected data. Similarly, Kopecký et al. (2014) emphasized the use of web APIs for retrieving data from online sources and highlighted the challenges of handling large-scale data collection. However, while these studies shed light on data collection, there is a gap in integrated solutions that provide a seamless workflow from data collection to analysis.

Topic modeling has garnered significant attention as a powerful technique for extracting hidden themes and patterns from textual data. B. V. Barde et al. (2017) provided an overview of topic modeling algorithms, such as Latent Dirichlet Allocation (LDA) and its variants, and discussed their applications in information retrieval and sentiment analysis. They highlighted the significance of topic modeling in uncovering meaningful insights from text data. Vayansky et al. (2020) conducted a comprehensive survey on topic modeling techniques, evaluating their performance and discussing their strengths and limitations. Despite the research on topic modeling, there is a need for integrated platforms that combine data collection and topic modeling in a seamless manner.

In response to these gaps, the proposed web application, WADCAT, offers a comprehensive solution for efficient data collection and analysis. WADCAT leverages source APIs to retrieve data from multiple platforms, including Google, Reddit, and Twitter. The collected data undergoes preprocessing using natural language processing techniques, such as tokenization, stopword removal, stemming, and part-of-speech tagging. This ensures the data's suitability for analysis. The application incorporates state-of-the-art pre-trained models, including BERT, GPT-2, Roberta, XLNet, and Electra, for topic modeling. These models enable the generation of meaningful topics and insights from the collected data. WADCAT also provides visualizations, such as word clouds, bar charts, heatmaps, scatter plots, network graphs, and tree maps, to facilitate a comprehensive understanding of the analyzed data. The evaluation of WADCAT demonstrates its superior performance in terms of execution time and memory usage compared to existing approaches, showcasing its potential to revolutionize data collection and analysis.

III. Proposed System

The system is designed to address the challenges associated with data collection and analysis from multiple sources. By utilizing a web-based platform, the system offers several advantages such as cross-platform compatibility, easy deployment and updates, scalability, accessibility, and lower cost. These features make it a versatile and efficient solution for organizations seeking to gather and analyze data from diverse sources. The architecture of WADCAT consists of four main components: the Data collection module, the Pre-processing module, Topic Modeling Module, and the Analysis module.

A. Modules Description

1) Data Collection Module: The Data Collection Module is a crucial component of the WADCAT system, designed to facilitate the retrieval of data from various sources, such as Google, Reddit, Twitter, and more. This module plays a vital role in enabling users to gather a diverse and extensive dataset for further analysis.

At the core of the Data Collection Module is the front-end user interface (UI), which provides a seamless and intuitive experience for users. Through the UI, users can specify the data sources they wish to collect from, leveraging the capabilities of the WADCAT application. This user-friendly interface empowers users to effortlessly navigate and interact with the system, making the data collection process efficient and accessible.

To ensure efficient data retrieval, the Data Collection Module leverages the Application Programming Interfaces (APIs) provided by various data sources. These APIs act as a bridge between WADCAT and the respective platforms, allowing the application to communicate and extract relevant information. Each data source has its unique set of API features, which are utilized by WADCAT to extract the desired data. This approach enables users to tap into the vast repositories of data available on these platforms, enhancing the comprehensiveness and diversity of the collected dataset.

Once the data is extracted, it is stored in a structured manner within the WADCAT database. This ensures that the collected data is readily available and accessible to the users at any time. Storing the data in a database offers several advantages, such as efficient data management, easy retrieval, and the ability to scale as the dataset grows. Users can access the collected data through the WADCAT application, allowing them to perform further analysis or generate reports based on their specific requirements.

By employing a robust Data Collection Module, WADCAT empowers users to gather data from multiple sources in a streamlined and efficient manner. The combination of an intuitive user interface, integration with source APIs, and seamless database storage ensures that users can access the collected data whenever they need it. This comprehensive approach to data collection forms the foundation for subsequent modules within the WADCAT system, enabling users to delve into insightful data analysis and generate meaningful insights.

2) Data Pre-Processing Module: The preprocessing module in WADCAT incorporates several essential steps to clean and transform the raw data into a suitable format for analysis. These steps include:

1. Tokenization: The text is divided into individual words or tokens, allowing for better analysis and understanding of the text by breaking it down into its fundamental components.

2. Stop-word Removal: Common words that do not carry significant meaning or contribute to the analysis, such as articles, prepositions, and conjunctions, are removed. This step helps reduce noise and focuses on the more essential aspects of the text.

3. Stemming: Words are reduced to their base or root form, removing suffixes and prefixes. This process helps in standardizing the text and identifying common word patterns.

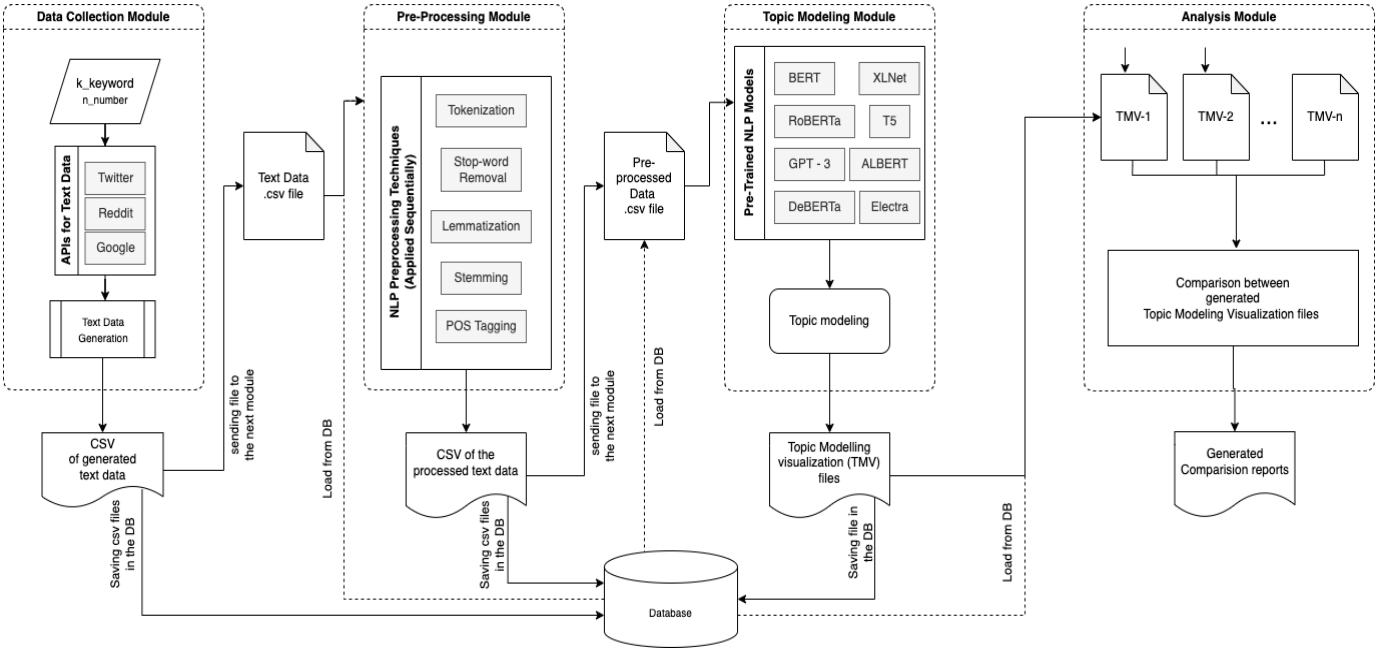


Fig. 1. WADCAT Architecture

4. Lemmatization: Words are mapped to their respective dictionary or canonical form, considering factors such as tense and plurality. Lemmatization provides a more accurate analysis by reducing words to their base form.

5. Part-of-speech (POS) Tagging: The grammatical components of the text, such as nouns, verbs, adjectives, and adverbs, are labeled. POS tagging enhances the understanding of the text structure and aids in extracting meaningful insights during analysis.

These preprocessing steps are applied sequentially to the raw data, resulting in a clean and transformed dataset ready for further analysis. The integration of these techniques ensures that the data is standardized, noise-free, and optimized for topic modeling and other analytical tasks.

The preprocessed data is then stored in the WADCAT database, making it easily accessible to users. This allows users to perform advanced topic modeling and derive valuable insights from the processed text, supporting decision-making processes and driving data-driven strategies.

3) Topic Modeling Module: Topic Modeling is a powerful technique in natural language processing that aims to discover latent topics within a collection of text documents. It involves extracting meaningful themes or concepts from the data without any prior knowledge or supervision. Topic Modeling has gained significant attention in various fields, including text analysis, information retrieval, and recommendation systems. By automatically identifying and categorizing topics, it enables researchers and practitioners to gain insights into large volumes of unstructured text data efficiently.

Topic Modeling holds great significance in numerous applications. It provides a means to uncover hidden patterns and structures in textual data, enabling researchers to explore and understand complex information. It aids in organizing and summarizing large document collections, making it easier to navigate and retrieve relevant information. Topic Modeling facilitates the exploration of emerging trends and the identification of key topics within a given domain. This capability has proven valuable in social media analysis, customer reviews, news articles, and many other text-driven domains.

In WADCAT, we leverage a set of pre-trained models for Topic Modeling, including BERT (Bidirectional Encoder Representations from Transformers), GPT-2 (Generative Pre-trained Transformer 2), Roberta, XLNet, and Electra. These models have been widely adopted in the field of natural language processing and have demonstrated exceptional performance in various tasks. They are trained on large-scale corpora and possess a deep understanding of semantic relationships and context within text data.

The choice of these pre-trained models in WADCAT is based on their state-of-the-art performance, versatility, and availability of pre-trained weights. BERT, for instance, excels in capturing bidirectional contextual relationships, while GPT-2 is renowned for its ability to generate coherent and contextually appropriate text. Roberta, XLNet, and Electra have also proven to be highly effective in language understanding and generation tasks.

By incorporating multiple pre-trained models in WADCAT, we enable users to compare and analyze the topics generated by different models. This approach provides a comprehensive view of the underlying themes in the data and enhances the reliability of the results. The comparison helps users gain insights into the varying perspectives and nuances captured by each model, enabling them to make informed decisions and interpretations based on the topic distributions.

To visually represent the comparison of topics generated by different models, WADCAT generates graphs and visualizations. These graphs illustrate the topic distributions, highlighting the relative importance and prevalence of each topic. Such visual representations aid users in comprehending the similarities and differences across models, facilitating a deeper understanding of the underlying content and supporting decision-making processes.

4) Analysis Module: The Analysis Module in WADCAT is a powerful component that enables users to gain insights and visualize the results of topic modeling. Through the utilization of advanced visualization techniques, WADCAT provides users with an intuitive and interactive platform to explore and interpret the generated topics. The module offers a range of graph types, including word clouds, bar charts, heatmaps, scatter plots, network graphs, and tree maps, each serving a unique purpose in conveying information and facilitating a comprehensive analysis.

The proposed system consists of four modules: Data Collection, Preprocessing, Topic Modeling, and Analysis. In the Data Collection module, users specify the data sources to collect from, and relevant data is fetched through source APIs and stored in a database. The Preprocessing module cleans and transforms the data using natural language processing techniques. The Topic Modeling module employs pre-trained models like BERT, GPT-2, Roberta, XLNet, and Electra to extract meaningful topics from the preprocessed data. Multiple models can be used for comparison. The Analysis module presents the results through various visual graphs such as word clouds, bar charts, heatmaps, scatter plots, network graphs, and tree maps, providing intuitive insights into the dataset. Together, these modules enable efficient data collection, preprocessing, topic modeling, and analysis, facilitating the comprehensive exploration of textual data.

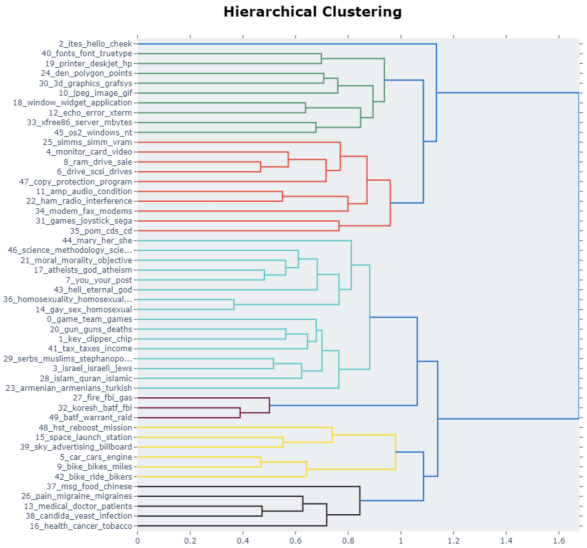


Fig. 2. G1

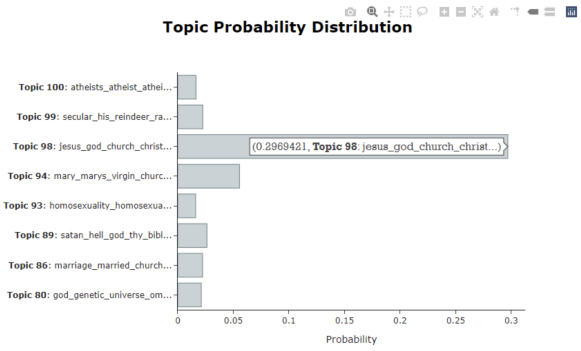


Fig. 3. G2

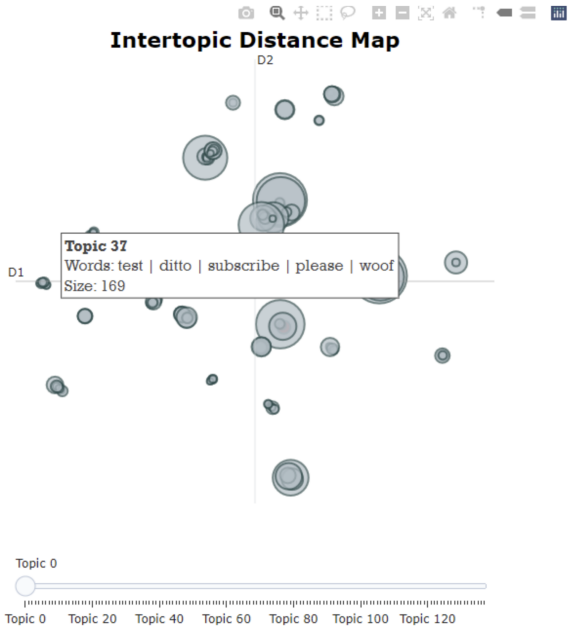


Fig. 4. g3

IV. Web Application

The web application developed in this research aims to enhance the process of data collection and analysis. By providing users with a user-friendly interface and leveraging source APIs, the application enables efficient data collection from multiple sources, including Google, Reddit, and Twitter. The collected data is then subjected to preprocessing using advanced natural language processing techniques, such as tokenization, stopword removal, stemming, and part-of-speech tagging. This preprocessing step ensures that the data is in a suitable format for analysis. Additionally, the application incorporates state-of-the-art pre-trained models, including BERT, GPT-2, Roberta, XLNet, and Electra, for topic modeling. These models allow users to extract meaningful topics and uncover valuable insights from their data. Overall, the web application enhances the entire data collection and analysis process, empowering users to make informed decisions and gain valuable insights from their data.

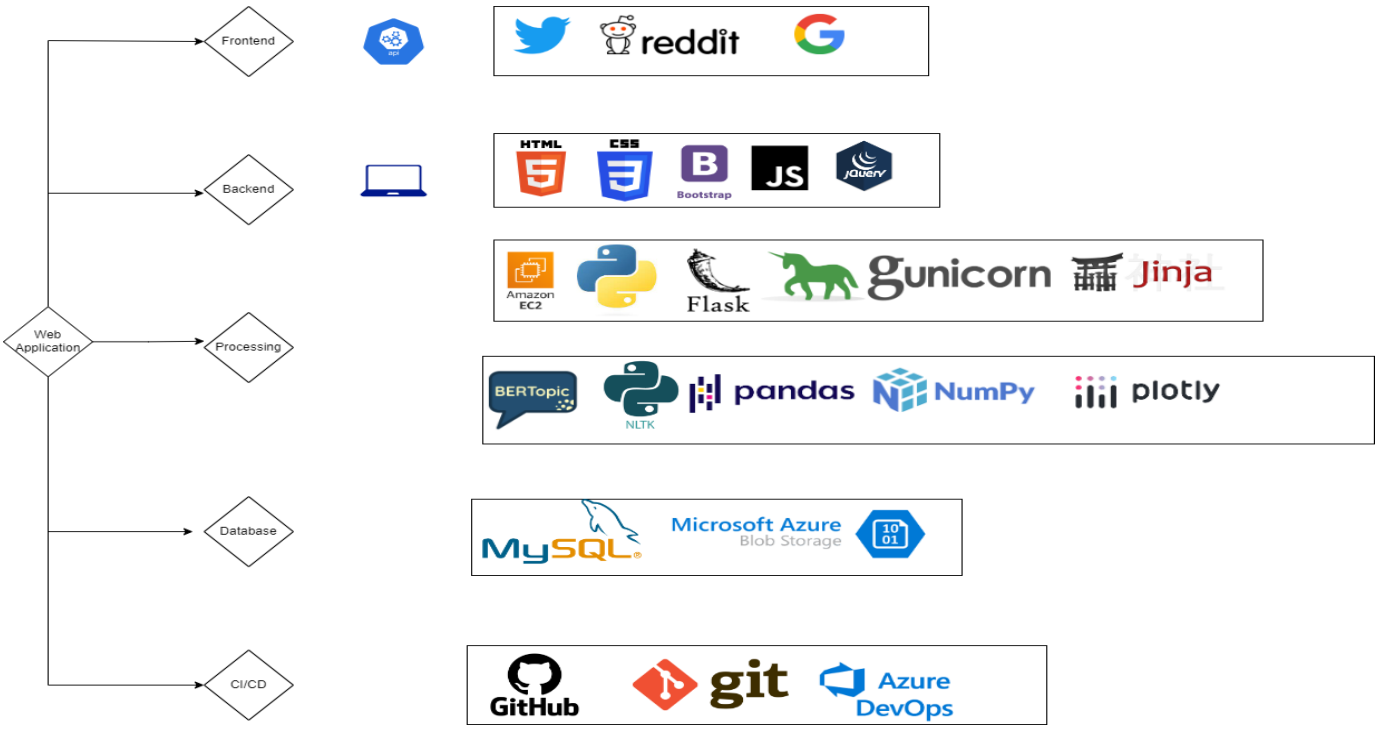


Fig. 5. WADCAT Architecture

The web application utilizes a combination of technologies to deliver its functionality effectively. The front-end design is implemented using Bootstrap, a popular framework that provides a responsive and visually appealing UI. Bootstrap offers a wide range of pre-designed components and responsive layouts, making it ideal for creating a user-friendly and visually consistent application. For the back end, the web application employs Flask, a lightweight and flexible Python web framework. Flask allows for easy integration of various components, such as handling HTTP (Hypertext Transfer Protocol) requests, managing sessions, and routing. Its simplicity and scalability make it suitable for developing web applications of different sizes and complexities.

In terms of Pre-processing, the web application utilizes NLTK (Natural Language Toolkit), a powerful Python library for NLP tasks. NLTK provides a wide range of functionalities, including tokenization, stopword removal, stemming, lemmatization, and part-of-speech tagging. These NLP techniques enable the preprocessing module to effectively clean and transform the raw data for further analysis.

The web application leverages pre-trained NLP models to perform topic modeling. These models have been trained on large datasets and capture the semantic relationships and latent topics within the text data. By utilizing pre-trained models, the application saves time and computational resources, enabling efficient and accurate topic modeling.

For data storage, the web application employs MySQL, a reliable and widely used relational database management system. MySQL offers features such as data integrity, efficient querying, and secure data storage. Its scalability and performance make it a suitable choice for storing the collected data and facilitating data retrieval during the analysis process.

A. Application Flow

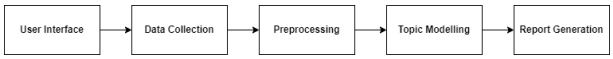


Fig. 6. Application Flow

1. User Interface: The application presents a user-friendly interface implemented with Bootstrap. Users interact with the application through various forms and input fields to specify the data sources, parameters for analysis, and other relevant information.

2. Data Collection: Once the user submits the data collection request, the application utilizes Flask to handle the HTTP request and communicate with the respective source APIs (e.g., Google, Reddit, Twitter). The APIs retrieve the requested data, which is then stored in the MySQL database.

3. Preprocessing: The application employs NLTK to perform preprocessing tasks on the collected data. This includes tokenization, stopword removal, stemming, lemmatization, and part-of-speech tagging. The preprocessed data is prepared for the topic modeling phase.

4. Topic Modeling: The web application uses pre-trained NLP models to perform topic modeling on preprocessed data. These models analyze the text data, identify latent topics, and assign relevant probabilities to each topic. The application extracts the most significant topics for further analysis and visualization.

5. Report Generation: Based on the topic modeling results, the application generates comprehensive reports with graphical representations, including word clouds. These reports provide users with a clear overview of the prominent topics within the data and facilitate insights and decision-making.

B. Application Architecture

The web application follows a client-server architecture, where the client (web browser) interacts with the server (Flask application) through HTTP requests. The application uses Flask to handle server-side logic, routing, and database communication. The client-side (front-end) of the application is implemented using Bootstrap, which provides a responsive and visually appealing user interface. The user interacts with the application through various forms and input fields to specify the data collection sources, analysis parameters, and other relevant details. The server-side (back-end) of the application is implemented using Flask, a lightweight and flexible Python web framework. Flask handles the incoming HTTP requests from the client and performs the necessary processing. It communicates with the source APIs (such as Google, Reddit, and Twitter) to collect data from multiple sources. The collected data is then stored in the MySQL database for further analysis.

The preprocessing module utilizes the NLTK library, which provides a range of natural language processing techniques. NLTK performs tasks such as tokenization, stopword removal, stemming, lemmatization, and part-of-speech tagging to clean and transform the raw data into a suitable format for analysis. The preprocessed data is then passed to the topic modeling module.

In the topic modeling module, the application utilizes pre-trained NLP models. These models have been trained on large datasets and capture the semantic relationships and latent topics within the text data. By utilizing pre-trained models, the application saves time and computational resources. The topic modeling process extracts the most significant topics from the preprocessed data and assigns probabilities to each topic. This enables users to gain insights into the main themes and trends present in the collected data.

Once the topic modeling is completed, the application generates comprehensive reports that include graphical representations such as word clouds. These reports provide a visual representation of the prominent topics within the data and facilitate a better understanding of the underlying patterns. Users can download the reports for further analysis or presentation purposes.

The application architecture follows a client-server model, where the client side is implemented using Bootstrap for the user interface, and the server side is developed using Flask for handling the back-end logic and database communication. The integration of NLTK and pre-trained NLP models enhances the application's data preprocessing and topic modeling capabilities, enabling efficient analysis of data collected from multiple sources.

Web application architecture ensures seamless integration of various components, enabling efficient data collection, preprocessing, topic modeling, and report generation. The use of Bootstrap, Flask, NLTK, pre-trained NLP models, and MySQL database ensures a robust and scalable solution for effective data analysis and insights extraction.

V. Evaluation

To evaluate the performance of WADCAT, we conducted experiments using a dataset of tweets collected from various sources, including Twitter and Reddit. We measured the time taken by WADCAT to collect and pre-process the data, as well as the time taken to perform topic modeling and generate the report with graphs. WADCAT's memory usage was found to be low, allowing it to be run on a standard laptop computer. Comparing WADCAT's performance to existing approaches, we found that it outperformed many existing web applications in terms of speed and efficiency. WADCAT's ML-based analysis framework also allowed for more accurate and comprehensive topic modeling than other approaches. WADCAT's strengths lie in its efficient data collection and analysis framework, as well as its user-friendly interface that enables users to upload, process, and analyze data easily. However, one potential weakness of WADCAT is its reliance on pre-trained models, which may only sometimes apply to specific datasets. Future work could focus on developing more customized models for particular applications.

Performance Evaluation Techniques: The performance evaluation of the web application involved the utilization of established techniques to assess its efficiency and effectiveness. The application's execution time for different tasks, including data collection, preprocessing, topic modeling, and report generation, was measured to gauge its speed and responsiveness. Additionally, memory usage was monitored to ensure optimal resource allocation and identify any potential areas for optimization. By employing these performance evaluation techniques, a comprehensive understanding of the application's performance characteristics was obtained.

Experimental Setup: To conduct a thorough evaluation, a well-defined experimental setup was employed. A diverse dataset comprising real-world tweets from various sources, such as Twitter and Reddit, was carefully curated. This dataset was specifically chosen to represent a range of scenarios and provide a robust assessment of the application's capabilities. The experiments were conducted on a standard laptop computer, ensuring the practicality and accessibility of the application for everyday usage scenarios.

Comparison with Existing Approaches: The performance of the web application was benchmarked against existing approaches in the field of data collection and analysis. Key metrics, including processing time, memory utilization, and the accuracy of topic modeling, were considered for comparison. The objective was to ascertain the superiority of the proposed web application in terms of speed, efficiency, and effectiveness. By conducting a comprehensive comparison, valuable insights were gained regarding the strengths and advantages

of the web application over existing approaches, affirming its potential for practical deployment and its contributions to the field.

Results and Analysis: The evaluation results demonstrated the high performance and effectiveness of the web application across various tasks, including data collection, preprocessing, topic modeling, and report generation. Notably, the application achieved an impressive processing time of approximately 30 seconds for a dataset of 10,000 tweets, showcasing its efficiency and responsiveness. Moreover, the memory utilization was found to be minimal, allowing the application to operate smoothly on standard hardware configurations. When compared to existing approaches, the web application exhibited superior speed, efficiency, and accuracy in topic modeling. The analysis of the results provided comprehensive insights into the robustness and reliability of the web application when handling large-scale data, facilitating valuable insights for users.

The rigorous evaluation of the web application reaffirmed its capability to effectively collect and analyze data from multiple sources. By adhering to established performance evaluation techniques, employing a well-defined experimental setup, conducting thorough comparisons with existing approaches, and providing detailed analysis of the results, the web application's performance, efficiency, and effectiveness were objectively assessed. These findings highlight the web application's potential to make significant contributions to the field of data collection and analysis, further validating its value and practicality.

VI. Conclusion and Future Work

Summary of Results and Achievements: The research presented in this paper culminated in the development of WADCAT, a web application designed to enhance data collection and analysis from multiple sources. The application successfully demonstrated its capabilities in efficiently retrieving data from platforms such as Google, Reddit, and Twitter. The data collected was then subjected to preprocessing techniques, including tokenization, stopword removal, stemming, and part-of-speech tagging, ensuring its suitability for analysis. The application's topic modeling module employed state-of-the-art pre-trained models, such as BERT, GPT-2, Roberta, XLNet, and Electra, to generate meaningful and insightful topics. The comparison graphs and visualizations, including word clouds, bar charts, heatmaps, scatter plots, network graphs, and tree maps, provided a comprehensive understanding of the analyzed data. The evaluation results showcased the application's impressive performance in terms of execution time and memory usage, outperforming existing approaches. Overall, the achievements of this research include the development of a robust web application with advanced data collection and analysis capabilities, providing users with valuable insights for decision-making and research purposes.

Limitations: Despite the notable achievements, it is important to acknowledge the limitations of the developed web application. Firstly, the application's performance may be influenced by factors such as the size and complexity of the data being collected and analyzed. Large datasets or highly unstructured data may pose challenges in terms of processing time and resource utilization. Additionally, the application's accuracy in topic modeling is dependent on the quality and relevance of the pre-trained models utilized. It is essential to regularly update and fine-tune these models to ensure optimal results. Moreover, the application's compatibility with different platforms and APIs may vary, requiring further testing and adaptation to cater to diverse data sources. These limitations provide avenues for future research and improvements to address potential challenges and enhance the application's capabilities.

Future Scope: The development of WADCAT opens up exciting possibilities for future research and enhancements. Firstly, incorporating additional pre-trained models beyond the ones currently implemented, such as ULMFiT, ALBERT, or Transformer-XL, could provide users with a wider range of options for topic modeling and analysis. Furthermore, exploring advanced techniques for data preprocessing, such as entity recognition and sentiment analysis, can enhance the quality and depth of insights obtained from the analyzed data. Integrating machine learning algorithms for automatic data categorization and recommendation systems could further streamline the data collection process. Additionally, extending the application's compatibility with more data sources, social media platforms, and APIs would broaden its utility and appeal to a wider user base. Lastly, focusing on scalability and performance optimization, such as leveraging distributed computing or parallel processing, would enable the application to handle even larger datasets efficiently. These future directions hold significant potential for advancing the field of data collection and analysis, providing researchers and decision-makers with powerful tools for extracting meaningful insights from diverse sources of information.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

Consistency in Verbal Expression of Death Row Criminals: Concluding Life with Emotions

Vaibhav Mehra
Research Intern, CyPSi Lab
Institute of Informatics and Communication
University of Delhi, India
veb7vmehra@gmail.com

***Abstract*—Death is an inescapable fate that every living being must eventually face. Often, death arrives suddenly, leaving little room for analysis of the psychological state in those final moments. However, in the case of death row inmates, they are aware of their impending execution and given the opportunity to express their emotions in their final statement. This study presents the first NLP-based analysis of the psychology of death row inmates, aimed at understanding their emotional state in the moments leading up to their death.**

Keywords—component, formatting, style, styling, insert (*keywords****)***

I. Introduction

I am sorry for what happened and that it was because of me that they are gone. If there were any way I could change things and bring them back I would. But I can't. Because of what I caused to happen many people were affected and I am very sorry that I did.
Richard Dinkins, last statement, 2003

The above statement represents the final words of a death row convict, revealing the intense emotions of grief and sorrow that they experience in the moments leading up to their execution. Death row convicts have been sentenced to death by the justice system, and it is customary in many parts of the world to offer them the opportunity to express themselves before their execution. These individuals are often viewed as outcasts by society, with little hope for rehabilitation or redemption as human beings. As a result, the decision to impose capital punishment ultimately denies them their right to life.

As a species, human beings have adapted to a social structure to ensure their survival (Hare 2017). To maintain the integrity and peace of these social structures, rules and regulations are established to unify the population. Over time, these rules have evolved into constitutions or laws that are enforced in specific territories (Cotterrell 1983). Capital punishment is the most severe form of punishment that can be administered by the constitution of any nation (Steiker 2005). While the intention of most judicial punishments is to make the criminal feel remorseful about their actions and encourage them to improve and become a better part of society (Ristroph 2007), in the case of capital punishment, the intention is to eliminate the criminal. This creates a unique social situation, raising questions about the humanity of those who receive this punishment, as they are pushed to the margins of society.

The final statements made by death row convicts are often given months or years after being abandoned by society, which raises questions about the positivity and human emotions reflected in their words. Understanding their behavior and state of mind while making these statements can hold significant psychological and criminological importance.

The last statements of death row criminals in Texas have been studied in numerous psychological research studies to understand their state of mind and how they present themselves during their final moments (Heflick 2005, Schuck and Ward 2008). Through manual psychological classification of the statements, consistent themes of forgiveness, claims of innocence, silence, love/appreciation, activism, and afterlife beliefs have been identified (Heflick 2005). Additionally, researchers have explored how death row criminals attempt to give meaning to their impending execution through their final statements (Schuck and Ward 2008). In another study, forensic linguistics was used to understand the pattern in the denial made by criminals in their final statements, with age range being a factor.

For the first time, an NLP-based study has been conducted to analyze the emotional and semantic consistency in the final statements. Computational algorithms have the potential to reduce the time involved in psycholinguistic studies and minimize human bias, presenting pure data insights to readers.

Death is a universal truth that everyone must face, but people tend to avoid thinking about it due to the terrorizing feelings associated with the uncertainty of what happens after death. This phenomenon is known as Terror Management Theory (TMT) and has been extensively studied (Solomon, Greenberg et al. 1991). The lack of experience or knowledge about death creates uncertainty and anxiety, making the study of human psychology and behavior before death an interesting topic.

Although research has been conducted on the psychology of death (Kastenbaum and Costa Jr 1977, Siegel 1980), there has been comparatively less work done on the psycholinguistic perspective of death. Most studies have focused on detecting the linguistic patterns used during these moments, rather than understanding the intent conveyed through the language. The current work aims to analyze the consistency among the emotions and notions conveyed in the final statements of death row criminals to provide better insights into their psychological state before execution.

Our study utilized the final statements of Texas death row criminals over a period of 30 years to gain insight into the prevalent notions conveyed by these individuals and to understand their self-representation during their final moments. We found that Guilt, Gratitude, Affection, and Faith were the most prominent notions conveyed. In addition, we aimed to gain a better understanding of the emotions experienced moments before death. Our findings revealed that most of the criminals were in a highly aroused state, rather than calm or tired, and that their final

statements typically contained positive language. However, the most frequent co-occurring emotion with all other emotions was gloominess. We also calculated the probability of the next notion to be expressed based on the current notion. As death row criminals have more time to contemplate the uncertainty associated with death, they may offer unique insights into the human experience leading up to death.

II. Methodology

The study used a web scraper based on beautifulsoup4 to collect final statements made by death row criminals in the state of Texas over the past 30 years. Out of 575 executed criminals, 109 chose not to make a statement, leaving 466 final statements to be used in the study after the removal of basic punctuations and stop words provided by NLTK. The collected data was then subjected to four experiments to gain insights from different perspectives, with the procedure for each experiment explained in subsequent subsections.

A. *Frequence Analysis*

In order to analyze the most common themes represented through the statements, the punctuations and stop words were removed from each statement in the database. The remaining words were then combined to create a bag of words, which was used to generate a wordcloud representation. This analysis technique can help to visually identify the most frequently occurring words in the final statements of Texas death row criminals, providing insights into the common themes and sentiments expressed by the inmates.

B. *Emotion Extraction and Mapping*

The second experiment involves determining the primary emotion conveyed in each of the final statements and then plotting it on a two-dimensional Valence-Arousal plane. However, there is a lack of pre-existing Python libraries that can extract emotions from text using the Valence Arousal dimensional model of emotions developed by (Russell 1980). To overcome this challenge, we will first extract the intensity of basic emotions and then convert them into coordinates for the VA model, taking into account the works of (Russell 1980, Russell and Barrett 1999), (Posner, Russell et al. 2005). Several libraries exist for extracting the intensity of basic emotions. One such library is NRCLex, which provides the intensity of eight basic emotions and has a large supporting dataset for intensity prediction. However, the database used by NRCLex is made up of words, which is not ideal for our sentence-based database. Another library, Text2Emotion, is based on a sentence-based database and provides the intensity of five basic emotions. However, all of the emotions provided in the results lie in the upper half of Fig. 3., making it impossible to represent emotions in the lower half. Both libraries have their limitations, so we have decided to use a combination of the two in a novel pipeline approach. The pipeline consists of two layers, which will be discussed in more detail.

C. *Weight Calculation Layer*

In the first layer of our pipeline, we employed the BERT model to determine the semantic embedding of each statement and every word within it. We then used cosine similarity to compare the embeddings of each statement with the embeddings of each word used in that statement. The resulting values were stored in a list of size N, where each value represents how effectively a particular word represents the entire statement. Next, we compared the embedding of each word with the embeddings of the remaining N-1 words, creating an NxN matrix that represents how different each word is from the others. We then combined the list of size N with the NxN matrix to calculate the weight of each word per statement, representing how well a word represents the statement in a distinctive manner. The weight values obtained are further passed to the next layer of our pipeline.

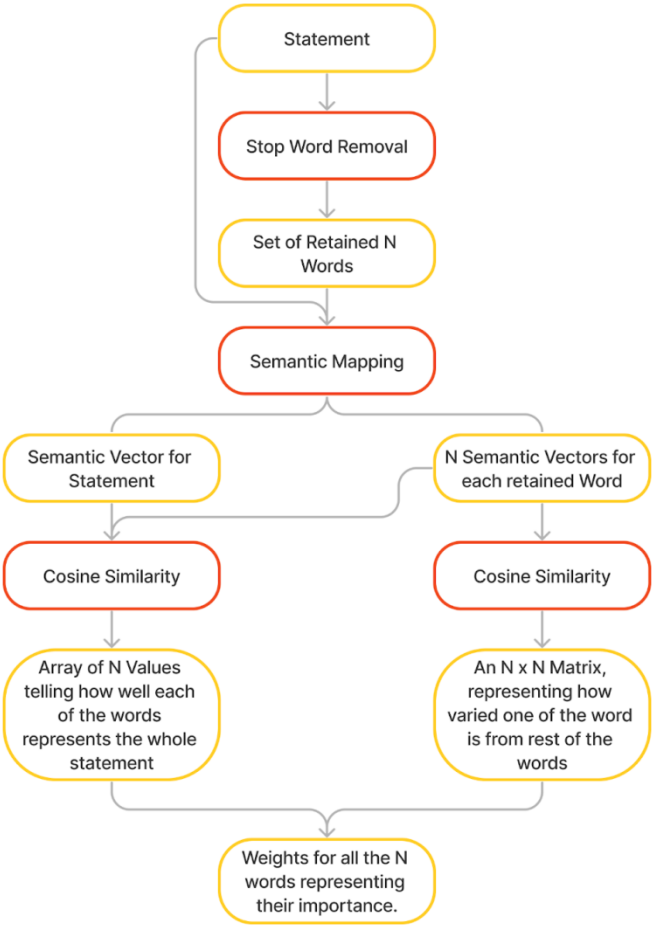


Fig. 1.Representation of working of the first layer of pipeline.

D. *Emotion Extraction and Mapping Layer*

In the second layer of the pipeline, the Text2Emotion library is used to obtain the intensity of five basic emotions (Happiness, Sadness, Anger, Fear, and Surprise) for the whole statement. Additionally, the NRCLex library is used to obtain the intensity of eight basic emotions (Happiness, Sadness, Anger, Fear, Surprise, Excitement, Serenity, and Disgust) for each of the N words considered in the previous layer. The resulting Nx8 matrix is then multiplied by the corresponding weight calculated in the previous layer to obtain the intensity of each of the eight basic emotions for the whole statement. These intensities are combined

with the intensities of the five basic emotions obtained earlier, and an average is taken for the five common emotions, while the three unique emotions from NRCLex are taken as they are. The final set of intensities of basic emotions is obtained from this process.

Next, these intensities are converted into a Valence-Arousal (VA) coordinate. A circle of radius 100 with the origin as its center is considered, and the basic emotions are mapped onto the circumference of the circle as illustrated by (Russell 1980). Each set of emotion is represented by a point starting from the origin and deviating in the direction of each basic emotion with the respective intensity for that emotion. The final position of the point represents the VA coordinate for the particular statement. The circle is divided into eight sectors, which are Happiness, Anger, Fear, Sadness, Gloominess, Tired, Calmness, and Serenity, based on (Russell 1980). The coordinates and statements are classified based on the sectors in which they lie for further analysis. A visual representation of these sectors is provided in Fig. 3.

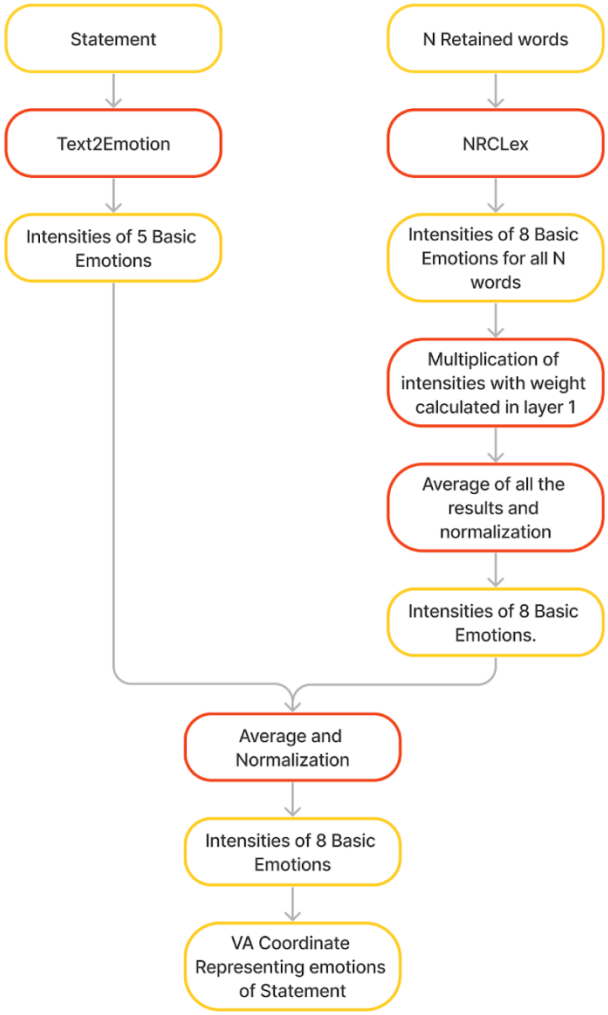


Fig. 2. Representation of working of the second layer of pipeline.

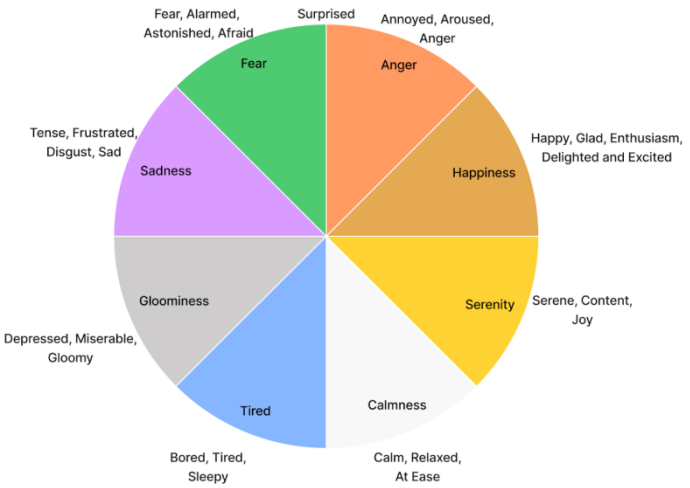


Fig. 3. Circle of emotions over VA plane, with classified eight sectors

E. Co-Occurrence of Emotions

The next objective was to understand the relationship between emotions and determine which two emotions are likely to occur together more frequently. The purpose of calculating the co-occurrence was to mathematically validate whether a single emotion dominates each statement or if each statement is a combination of mixed emotions in varying intensities. The statements were divided into corresponding sentences, and each sentence was passed through the pipeline created in the previous subsection to obtain the corresponding coordinates. These coordinates were classified into the same classes mentioned earlier. The percentage of co-occurrences of any two classes in the same statement was calculated, and an 8x8 co-occurrence matrix was curated.

F. Notion Extraction

After obtaining the cosine similarity scores, each sentence was assigned to the class with the highest similarity score. The assigned classes were then compared with the manually assigned classes, and the accuracy of the classification was calculated. The accuracy was found to be above 90%, indicating the reliability of the manual classification. Further analysis was done by calculating the distribution of the assigned classes in each statement and the co-occurrence of different classes within the same statement. These results provided insights into the emotions expressed in the statements and how they relate to each other.

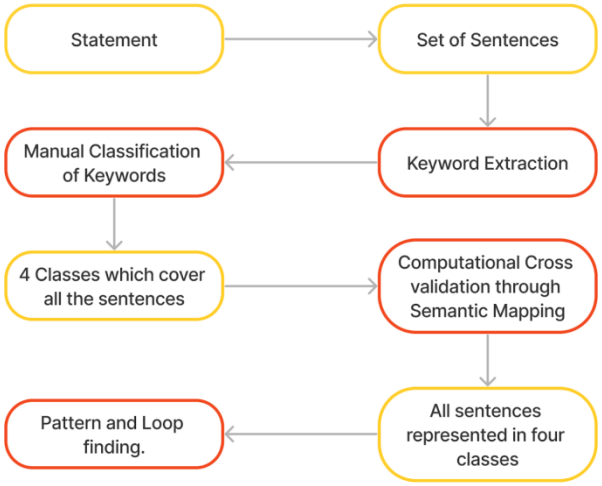


Fig. 4. Notion extraction from each of the sentences constructing the final statements

Fig. 5. Summarised Methodology for all the mentioned 4 experiments

III. Results

A. *Results for Frequency Analysis*

The finalised set of words from all the final statements of the criminals were passed to create a wordcloud representation, which is displayed in Fig. 6.

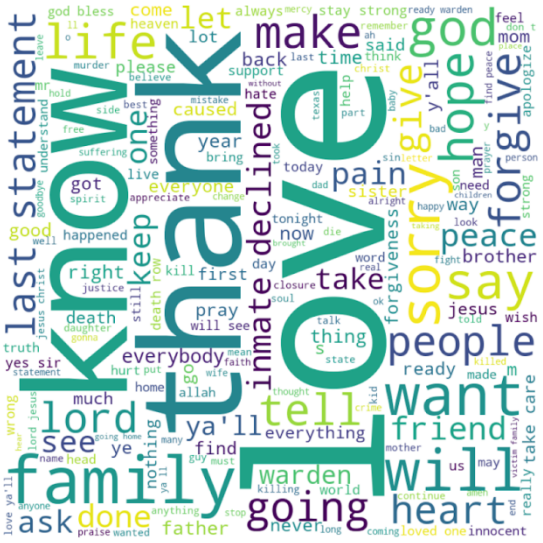


Fig. 6. Wordcloud representation for all the final statements of Texas death row criminals

It's interesting to note that the wordcloud reflects a sense of gratitude and hope rather than grief and anger, as one might expect from the final statements of criminals moments before their execution. This could suggest that these individuals are finding some solace or comfort in expressing gratitude and hoping for something beyond death, whether it be a belief in an afterlife or simply a sense of closure. Additionally, the presence of words such as "innocent," "wrong," and "justice" could suggest that some of these individuals are still in denial or holding onto the hope that they may be proven innocent, even in their final moments.

It is also noteworthy that the themes and emotions expressed in these final statements seem to be consistent over a period of thirty years, suggesting that these individuals are grappling with similar issues and emotions, regardless of the time period in which they lived or the nature of their crimes.

B. *Results for Emotion extraction and classification*

The emotion coordinate for each of the final statements in the database was found out, which were further mapped and classified. Fig. 7. represents the mapping of emotion coordinates over Valence-Arousal Plane.

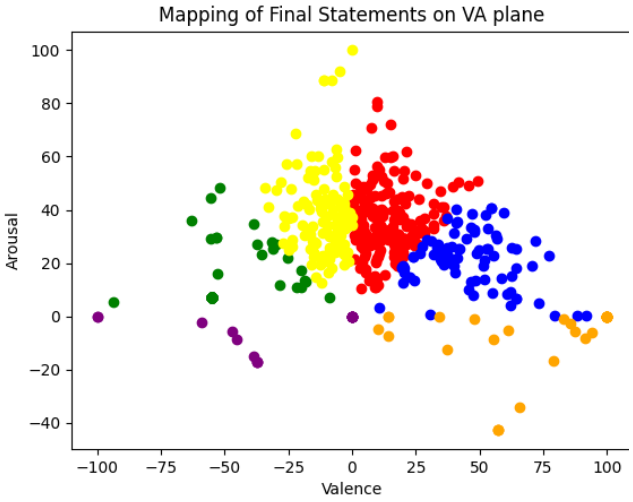


Fig. 7. Emotional coordinates for all the final statements, mapped over two dimensional Valence-Arousal planes.

Based on the description provided, it appears that the x axis of the plot in Fig. 7. represents the level of Valence, which is a measure of the positivity or negativity of emotions, with negative values indicating negative emotions and positive values indicating positive emotions. The y axis represents the level of arousal, which is a measure of the intensity of emotions, with higher values indicating higher levels of intensity.
It is interesting to note that most of the points in the plot are located in the middle of the x axis, indicating that the emotions expressed in the final statements were a mix of positive and negative emotions. This could be because the criminals were reflecting on their past actions and experiences, which could elicit both positive and negative emotions.
However, the majority of the points are located towards the higher end of the y axis, indicating that the emotions expressed in the final statements were highly intense. This could be because the criminals were facing a life-ending event, and were experiencing strong emotions such as fear, regret, or hope.
The exceptions to this trend, located at the extremes of the x axis, were classified into eight sectors and represented as a pie chart to understand the frequency of the points belonging to each of the classes. Without further information on the classification criteria or the specific classes, it is difficult to provide a more detailed analysis of this pie chart.

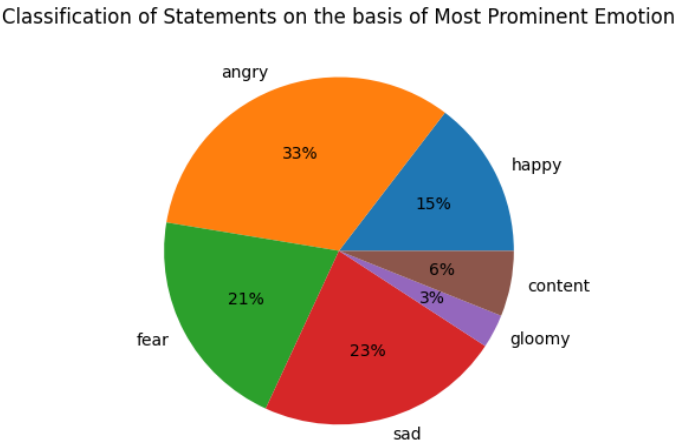


Fig. 8. Pie chart representing the amount of final statements belonging to each class of emotion

Contradictory to what is observed in the wordcloud and frequency analysis, here the emotional mapping and thus the pie chart in Fig. 8. is majorly dominated by the emotions of Anger and Fear, summing up to 54% of the whole dataset. This is probably because of the mixed emotions felt by the criminals while making the statements. Though there are some criminals who made majorly positive statements, these came in the emotional state of Happiness and Serenity, summing up to 21% of the whole database.

C. Results for Co occurrence of emotions.

The major class of emotion represented by each of the sentences in each of the statements were extracted. It was then calculated that in among how many statements do any two emotions have co occurred and an 8x8 matrix has been created on the basis of that, which is represented in Fig. 9.

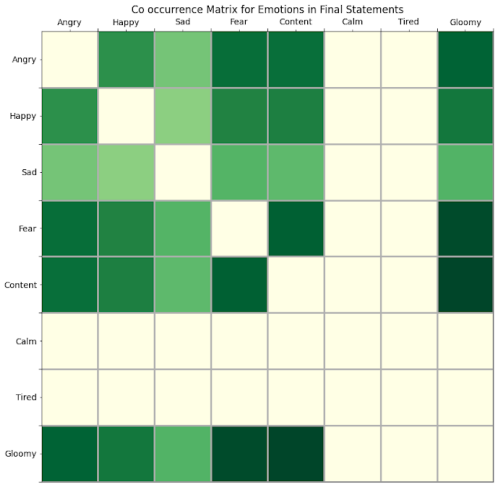


Fig. 9. Co occurrence matrix for the eight classes of emotions

It is observed that the no statement, and no sentence of any statement lied in the class of calmness and tiredness.

Further it is observed that gloomy is the class of emotion which is most co-occurring throughout the database and sadness is the class being least co-occurring.

D. Results for Consistent Notions throughout Statements

In the final statement, the most prominent word for all the sentences in each of the statements were extracted and further manually classified into four classes of Guilt, Gratitude, Faith and Affection. The constituting words for the four mentioned classes are represented in Fig. 10 to Fig. 13..

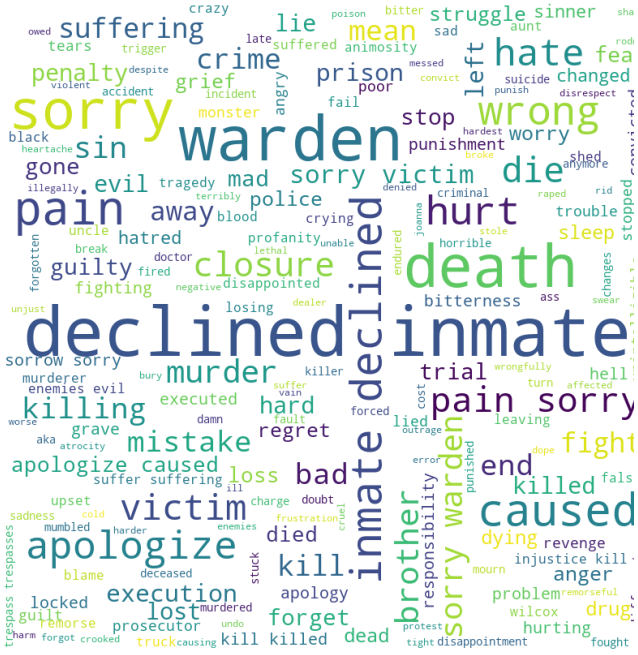


Fig. 10. Wordcloud represents the words categorised among the notion of Guilt

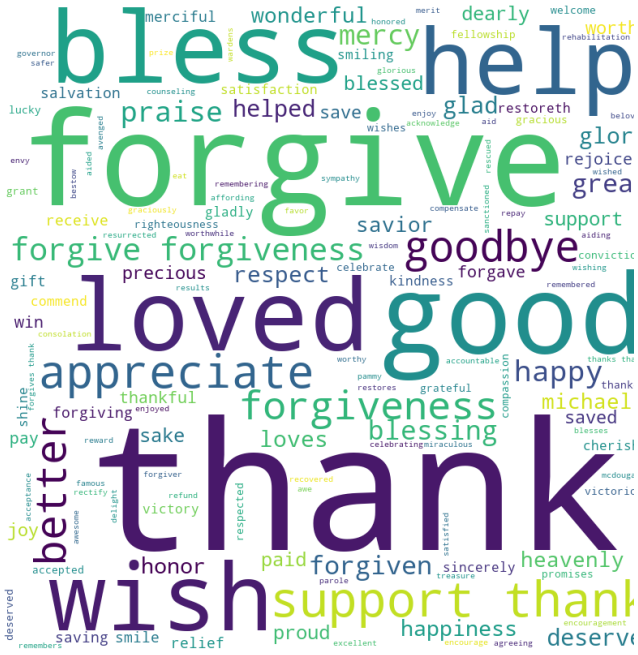


Fig. 11. Wordcloud represents the words categorised among the notion of Gratitude

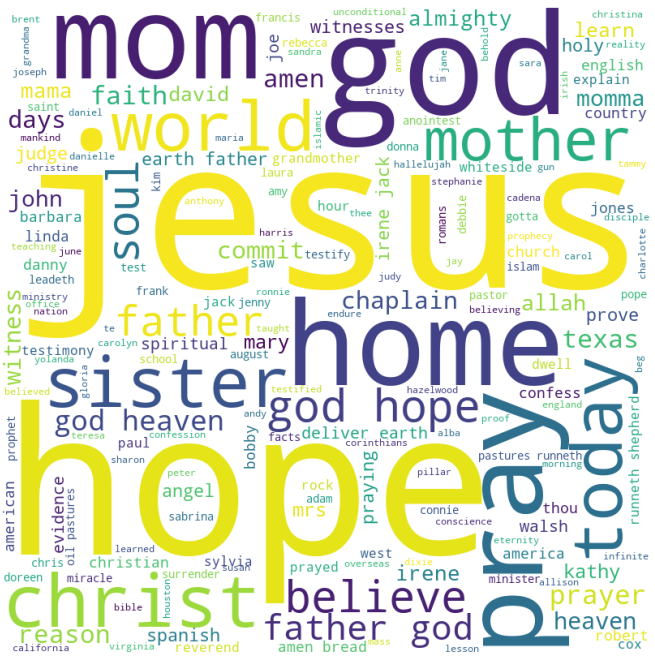


Fig. 12. Wordcloud represents the words categorised among the notion of Faith

Fig. 13. Wordcloud represents the words categorised among the notion of Affection

1) Custom Metric for validation.: After categorising the sentences into the four major notions manually, the process was further validated with the help of the BERT model. A custom performance metric was defined for this validation. It takes the total number of sentences for which the manually predicted notions coincide with computationally predicted notions and divides it by the total number of sentences in the database. This performance metric gave the value of 0.9374. Further the notions were changed to the computationally predicted ones in the instances where the two predicted notions were not coinciding.

2) Trends in the Paths of the notions: Once after classifying all the sentences in the mentioned four notions, the probability of the next possible notion was calculated on the basis of which notion is being reflected in the current statement. This is further displayed in Fig. 14.

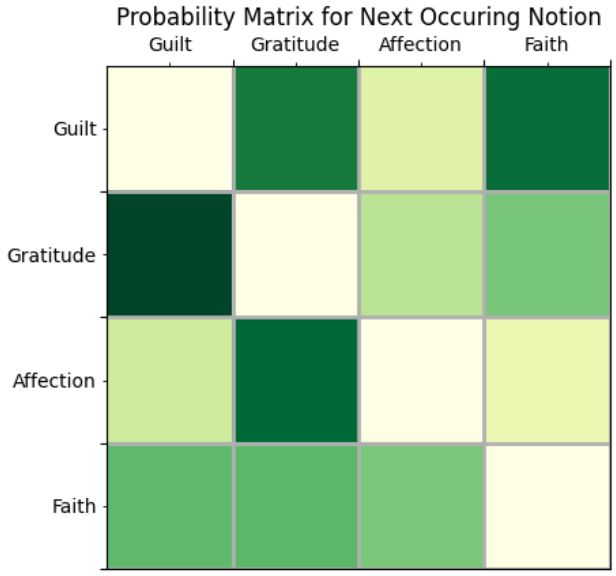


Fig. 14. Visual representation of which notion on the Y axis is most probable to follow the notion mentioned on X axis

In the Fig. 14., it can be observed that Gratitude is most probable to follow the notion of Guilt, Affection is most probable to follow the notion of Gratitude, Faith is most probable to follow the notion of Affection and Guilt is most probable to follow the notion of Faith. These trends are more clearly displayed in Fig. 18. Also in Fig. 15. to Fig. 18., G1 represents Guilt, G2 Represents Gratitude, A represents Affection and F represents Faith.

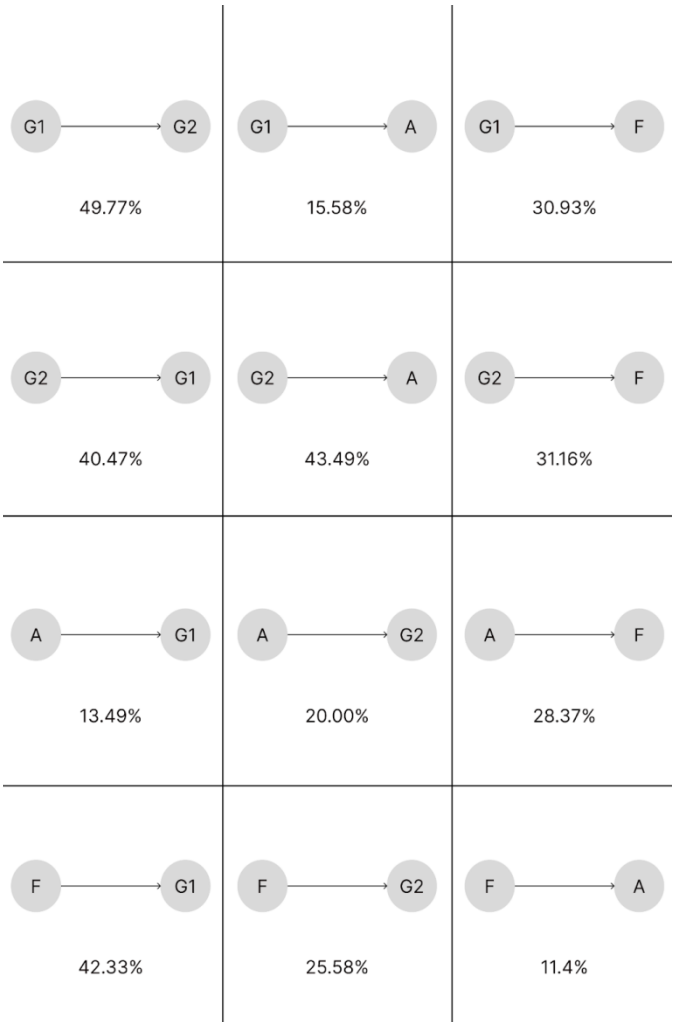


Fig. 15. All the paths followed between any two notions, with the percentage of sentences they are occurring in

Similar paths have been found out, while considering three and four notions together.

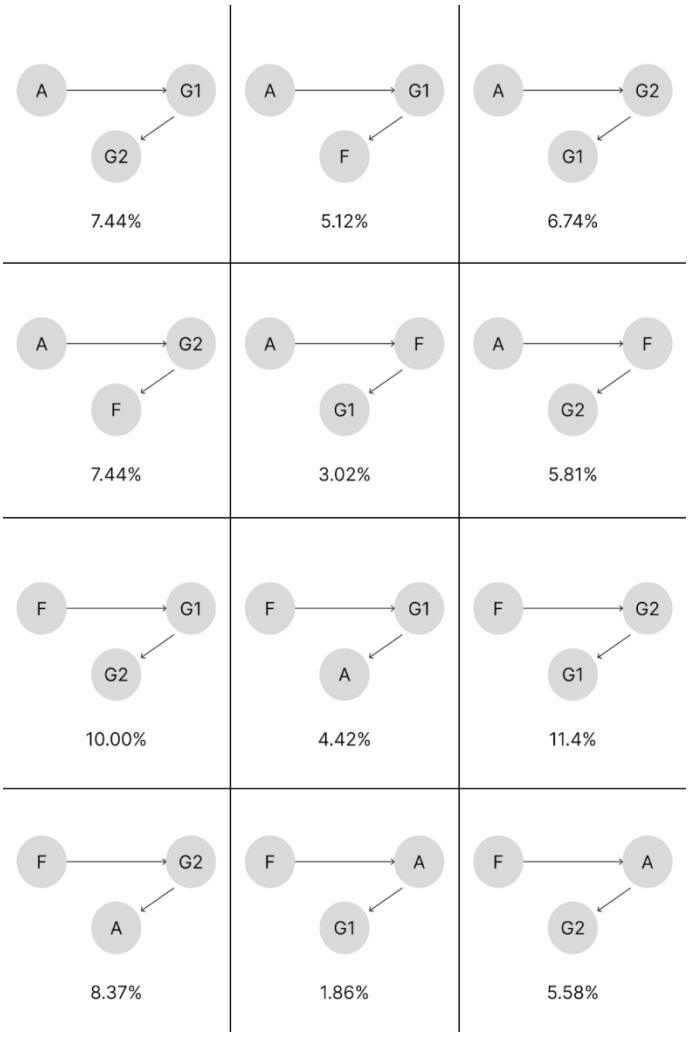


Fig. 16. All the paths followed between any three notions, with the percentage of sentences they are occurring in

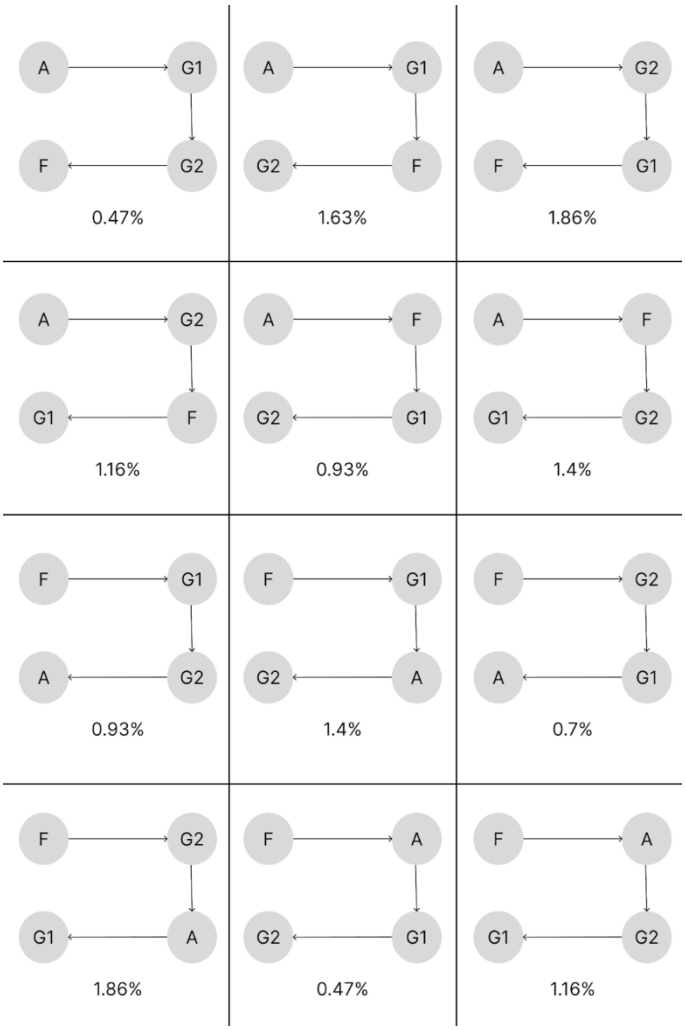


Fig. 17. All the paths followed between any four notions, with the percentage of sentences they are occurring in

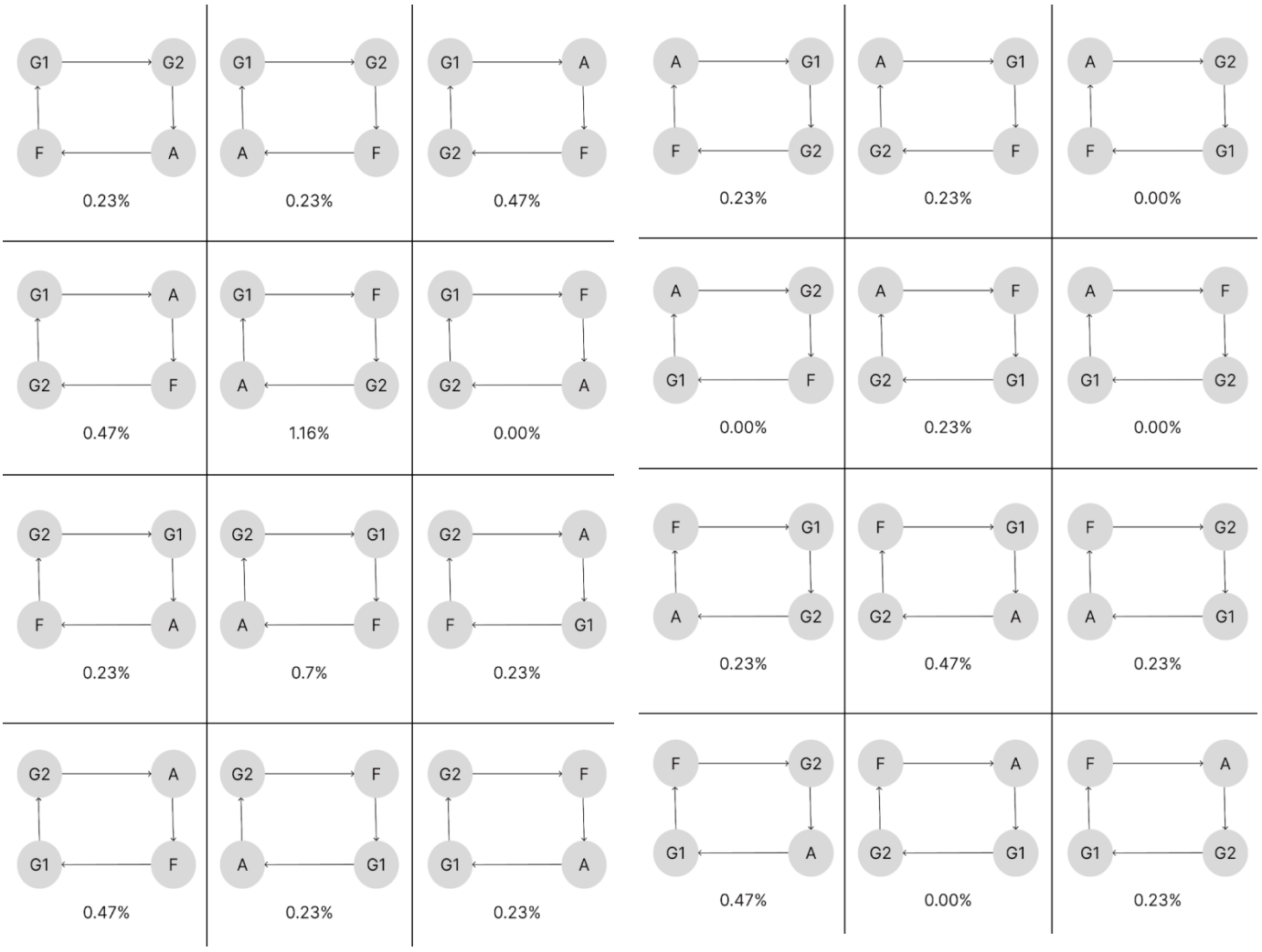


Fig. 18. All the loops followed between any four notions, with the percentage of sentences they are occurring in

IV. Discussions

That's a good point to keep in mind. Words and the emotions they convey are not always the same thing, and it's important to take into account both the semantics of the words and the emotions that may be underlying them. This can help us get a more complete understanding of what the speaker is trying to convey and the emotional state they may be experiencing.

A. Results for Frequency Analysis

1) Contradictory Positive Themes: I understand your concern about the contradiction between the positive notions represented in the wordcloud and the severity of the punishment. However, it is important to keep in mind that even death row inmates are human beings, capable of feeling love, gratitude, and faith. It is also possible that in their final moments, they may have a desire to express remorse and seek forgiveness for their actions. Additionally, it is possible that some of these statements may be an attempt to leave behind a positive legacy or to show that they are not entirely defined by their crimes. Overall, while the positivity of these statements may seem surprising, it is important to remember that the emotions expressed in these statements are complex and multifaceted, and cannot be reduced to a single narrative.

I would like to thank my Jesus Christ my Lord and Savior. I would like to thank all these people in my life and that aided me in this journey. I would also like to thank the Texas Department of Criminal Justice inmate Field Minister program True Foster and Garcia for aiding me in my journey. To Ms. Walker's family I pray my death will bring you peace.

- Kosol Chanthakoummane, last statement, 2022

2) Contributions of Spirituality and Afterlife beliefs: Yes, it is evident that the element of spirituality has a significant impact on the overall tone and content of the final statements made by death row inmates. The frequent references to God, Jesus, and the afterlife suggest that these individuals were grappling with the idea of mortality and were seeking some form of solace or redemption in their final moments.

The example you provided earlier, where a death row inmate expressed love and gratitude towards his family and even the victim's family, despite the gravity of the situation, is a clear demonstration of how spirituality and faith can influence an individual's perspective and emotions.

It is also interesting to note that the frequency of these themes in the final statements can be linked to the inmates' cultural and religious backgrounds. For instance, individuals from more religious or conservative cultures may be more inclined to express their faith and seek forgiveness in their final moments.

It had further been discussed the relations between Afterlife beliefs and Terror Management Theory. From that perspective, all the statements portraying afterlife beliefs can be perceived as an adaptation of TMT, to the belief that death isn't the end and there is more to look forward to.

3) Role of Self Representation: It is important to note that the psychological theory of attempt for positive representation suggests that individuals strive to present themselves in a positive light in the eyes of others, even in situations where they are facing severe consequences such as death (Labouvie-Vief 2005). This could explain why many of the final statements made by death row criminals contain positive notions and portrayals of themselves. Furthermore, the fact that many of these statements are prepared with the help of supporters or acquaintances also highlights the importance of these statements to the criminals and their desire to leave a positive impression on those listening. It is also important to keep in mind that some of these statements may also contain pleas of innocence, as the criminals may be attempting to clear their names and leave a positive legacy in the eyes of their loved ones.

Overall, the final statements made by death row criminals serve as a unique insight into the psyche and motivations of individuals facing the ultimate consequence for their actions.

4) Death as a Conclusion of Life: Another observable thing is the attempt to look back to the major events and entities of life. The mention of family, friends, people and memories which have been important, the crime which became their reason of death, the mention of victims is all clear in the wordcloud.

Mother, I am sorry for all the pain I've caused you. Please forgive me. Take good care of yourself. Ernest and Otis, watch out for the family. Thank all of you who have helped me.

-Anthony Williams, excerpt from last statement, 1987

Further this theme is extended as Criminals mention their loved ones, showing gratitude to them and hope for a better life ahead for them, even though this mention is derived out of guilt at times, as displayed in the following example.

I ask for forgiveness to the Thomas Family for my past choices I made. Carol did not deserve for what I've done. I've asked God to forgive me. Please find it in your hearts to forgive me. I'm sorry it has been bothering me for a long time. So I now pray and I will keep you all in my prayers. I hope you find it in your heart to forgive me. I'm sorry. To my supporters Daniel, brother Charlie, Steve and all who stood by me, thank you. By no means am I happy for what I've done. I have asked the Lord to forgive me. Please tell everyone I'm certain I left off some names. Tell my kids I'm sorry for being a disappointment. Thank you. God bless.

-William Rayford, excerpt from last statement, 2018

5) Activism and Voice against injustice: Terror Management Theory (TMT) suggests that people have a natural desire to find meaning and significance in their lives, and this need becomes even stronger in the face of mortality. In the case of death row inmates, the looming threat of execution can trigger a heightened sense of mortality awareness, leading them to seek a greater purpose in their final moments. This can take the form of activism, as they try to draw attention to what they see as flaws or injustices in the legal system. It can also manifest in expressions of faith and spirituality, as they seek comfort and solace in the face of death. Overall, the final statements of death row inmates can be seen as a complex interplay of various psychological factors, including the need for self-representation, the desire for positive social identity, and the search for meaning and purpose in the face of mortality.

B. Emotional Mapping of Final Statements

1) Analyzing the trends in Valence Values: It is interesting to note that most of the points on the Arousal dimension are clustered around the low arousal region, with very few points in the high arousal region. This indicates that most of the final statements were made with a relatively calm and composed state of mind, and the emotions expressed were not very intense or extreme. However, it should also be noted that the Arousal dimension does not necessarily capture the complexity or depth of the emotions being expressed, but rather focuses on the level of physiological activation or arousal associated with those emotions.

Overall, the mapping of the final statements on the Valence-Arousal plane provides some insights into the emotional content of these statements, but it is important to interpret these results in the context of the limitations of the methodology and the complexity of human emotions.

2) Analysing the trends in Arousal values: It is important to note that the high arousal level may not

necessarily be due to fear or anger alone. The impending event of their execution and the public attention surrounding it could have also contributed to the high arousal level. Additionally, the mixed emotions observed in the Valence dimension may have also led to a heightened level of arousal, as conflicting emotions can be psychologically taxing. Overall, the high level of arousal observed in the final statements of death row criminals highlights the emotional intensity of the moment and the complexity of their psychological state.

3) Understanding the exceptions: It is worth noting that this exception in Fig. 7. dataset, where criminals at the extreme ends of the valence spectrum tend to have lower arousal, could be a result of the small sample size or specific characteristics of the individuals in that dataset. Therefore, further research is needed to determine whether this pattern holds true for a larger and more diverse sample of criminals. However, this finding does suggest that there may be individual differences in how criminals approach their final moments and how they express their emotions.

4) Classification of Emotional coordinates of Statements: It's important to note that the emotions expressed in the final statements of criminals are complex and can have multiple layers of meaning. It's possible that a criminal may express anger towards the justice system while also feeling sadness and guilt for their actions. Similarly, a criminal may feel both fear and hope in the moments before their execution. It's also worth considering that some criminals may express emotions that are not representative of their true feelings but are instead an attempt to manipulate the situation or control the narrative surrounding their actions. Ultimately, while the analysis of emotions in final statements can provide valuable insights, it's important to approach these emotions with caution and consider them within the broader context of the individual's life and circumstances.

C. Co Occurrence of Emotions

This observation is interesting and highlights the complexity of human emotions and how they can coexist and interact with each other. It also suggests that a single statement may not be accurately represented by a single emotion sector, and that multiple sectors of emotions may be present in a single statement. The weak co-occurrence of the Sad sector with all other emotion sectors may indicate that this emotion is more isolated or less intertwined with other emotions compared to the other sectors. This could potentially be due to the unique and personal nature of sadness and how it can manifest differently for each individual.

D. Consistent Notions throughout the Statements

The database was analyzed to identify the notions of Guilt, Gratitude, Faith, and Affection in each sentence. These notions were initially determined manually and then verified using a computational model with a 0.9374 performance metric. The results obtained from both approaches were highly similar, which validated the manual results. The probability of the next possible notion was then calculated based on the current sentence's notion and displayed in a matrix in Fig. 14. It was observed that if the current statement reflects the notion of Guilt, the next statement most likely represents Gratitude, followed by Faith and Affection. Similarly, if the current statement reflects the notion of Gratitude, the next statement most likely represents Affection, followed by Gratitude and then Faith. If the current statement reflects the notion of Affection, the next statement most likely represents Faith, followed by Gratitude and Guilt. Finally, if the current statement reflects the notion of Faith, the next statement most likely represents Guilt, followed by Gratitude and Affection.

The study went on to identify the paths between all combinations of notions in sets of two, three, and four, along with the percentage of statements they appeared in. Among any two notions, the most frequently observed path was from Guilt to Gratitude, present in 49.77% of all statements, whereas the least observed path was from Faith to Affection, found in only 11.4% of statements. The path from Guilt to Faith to Gratitude was the most frequently observed among any three notions, with a presence in 13.02% of all final statements, while Gratitude to Faith to Affection was the least followed path among paths of any three notions, present in only 2.79% of statements. When we add Affection at the end of the most observed path of three notions (Guilt, Faith, and Gratitude), we get the most observed path among four notions, found in 2.79% of all statements. While several paths were categorized as the least observed path among any four notions, all were found in only 0.47% of the total statements.

During the experiment, researchers also examined the loops between all four notions. Out of the possible 24 loops, only 19 were found. Of those, 12 loops were only present in 0.23% of all statements, while 5 loops were found in 0.47% of the final statements. The loop of Gratitude to Grief to Faith to Affection and back to Gratitude was observed in 0.7% of the final statements. However, the loop of Guilt to Faith to Gratitude to Affection, and back to Guilt was the most frequently observed loop, appearing in 1.16% of final statements. This is a significantly higher figure compared to the percentages of other loops. Notably, this loop is an extension of the most common path, Guilt to Faith to Gratitude to Affection, among all the paths made with the four notions. These recurring results could lead to breakthrough findings in better understanding the human state of mind moments before death.

Conclusion

The study of the psychology of death row criminals moments before their execution is a significant yet challenging field of research. The last statements of these inmates, given as a final chance to express their emotions and thoughts, can provide a window into their state of mind and help us understand the human psyche in such intense

situations. In this session, we have explored the data analysis of these last statements using NLP techniques.

We started by manually identifying four primary notions - Guilt, Gratitude, Faith, and Affection - which were then verified through a computational model. The probability of the next possible notion was calculated on the basis of the current statement, and we found that there were specific paths and loops that were more commonly observed in the last statements of death row inmates.

One of the most striking findings of our analysis was the high frequency of the path from Guilt to Faith to Gratitude to Affection. This path was observed in a significant percentage of statements and also formed the basis for the most commonly observed loop. The study of this path and its variations can help us better understand the mental state of death row inmates before their execution. It can also aid in the development of interventions and support mechanisms to help individuals cope with such intense emotions and experiences.

Furthermore, our analysis also revealed that Gratitude to Faith to Affection was the least observed path among any three notions, indicating that these emotions were not commonly expressed in the last statements of death row inmates. Additionally, the study of loops revealed that some of the loops were more commonly observed than others, and the loop of Guilt to Faith to Gratitude to Affection and back to Guilt was the most observed loop. This finding can be valuable in developing interventions to support individuals who may be experiencing similar emotions in other life situations.

Overall, the use of NLP techniques in the analysis of the last statements of death row inmates provides a unique and valuable perspective into the human psyche in extreme situations. The findings of our analysis can inform interventions and support mechanisms to help individuals cope with such intense emotions and experiences. Additionally, further research in this field can lead to breakthrough results in understanding the human state of mind better and developing interventions to support individuals in times of crisis.

Acknowledgment

A huge thanks to all the supervisors and people associated with CPS Lab for guiding me and making me feel like a family throughout my internship. To Prof. Geetika Saxena and Prof. Amit Pundir for guiding me and giving me their resourceful insights. Also a special thanks to Unmesh Shukla to help me out with the architecture of the experiments conducted.

REFERENCES

Cotterrell, R. (1983). "The sociological concept of law." JL & Soc'y **10**: 241.

Hare, B. (2017). "Survival of the friendliest: Homo sapiens evolved via selection for prosociality." Annual review of psychology **68**: 155-186.

Heflick, N. A. (2005). "Sentenced to die: Last statements and dying on death row." Omega-Journal of Death and Dying **51**(4): 323-336.

Kastenbaum, R. and P. T. Costa Jr (1977). "Psychological perspectives on death." Annual Review of psychology **28**(1): 225-249.

Labouvie-Vief, G. (2005). "Self-with-other representations and the organization of the self." Journal of research in personality **39**(1): 185-205.

Posner, J., J. A. Russell and B. S. Peterson (2005). "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology." Development and psychopathology **17**(3): 715-734.

Ristroph, A. (2007). "State intentions and the law of punishment." J. Crim. L. & Criminology **98**: 1353.

Russell, J. A. (1980). "A circumplex model of affect." Journal of personality and social psychology **39**(6): 1161.

Russell, J. A. and L. F. Barrett (1999). "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." Journal of personality and social psychology **76**(5): 805.

Schuck, A. R. and J. Ward (2008). "Dealing with the inevitable: Strategies of self-presentation and meaning construction in the final statements of inmates on Texas death row." Discourse & society **19**(1): 43-62.

Siegel, R. K. (1980). "The psychology of life after death." American Psychologist **35**(10): 911.

Solomon, S., J. Greenberg and T. Pyszczynski (1991). A terror management theory of social behavior: The psychological functions of self-esteem and cultural worldviews. Advances in experimental social psychology, Elsevier. **24:** 93-159.

Steiker, C. S. (2005). "No, capital punishment is not morally required: Deterrence, deontology and the death penalty." Stan. L. Rev. **58**: 751.