International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 12, No. 3, 2025, pp. 493-511.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

An Optimized Machine Learning Approach for Scalable Online Public Opinion Mining

B. Balaji [a], Dr.D. Suresh [b]

[a] *Research Scholar, Department of Computer Science, St. Peter's Institute of Higher Education and Research, Chennai, India. E-mail: kpmbalaji.b@gmail.com, Orcid: https://orcid.org/0009-0001-1156-558X*
[b] *Associate Professor, Department of Computer Science, St. Peter's Institute of Higher Education and Research, Chennai, India. E-mail: sureshd.cs@spiher.ac.in, Orcid: https://orcid.org/0000-0003-3030-4294*

Abstract

The swift growth of the social media platforms resulted in tremendous amounts of unstructured textual data which means that online analysis of public opinion is a crucial activity in comprehending the perception and sentiment dynamics in the society. A good sentiment classification needs a good pre-processing, feature selection methods optimized to deal with noise, high dimensionality and scalability. This research introduces a unified model of online sentiment analysis of public opinion about something based on the Contextual-Adaptive Text Refinement Engine (CATR-E) to do text pre-processing, Federated Feature Ranking Protocol (FeFRP) to determine features, and Logistic Regression to classify sentiment. CATR is a systematic method of noise elimination emojis, URLs, and stop words and semantic relevance normalization and lemmatization. FeFRP identifies and chooses the most informative TF-IDF features which tremendously downsizes dimensionality, yet still captures important sentiment-carrying words. Logistic Regression is subsequently used to determine sentiments into predetermined categories according to the estimation of probabilities. The framework is effective as seen through experimental assessment on a real-world Twitter entity sentiment dataset with a Feature Selection 93.00 better than some of the current machine learning and deep learning Existing methods. The effectiveness of the framework is supported by its 93.00% accuracy of feature selection and 35.04% feature dimensionality reduction using the FeFRP protocol as well as 43.79% of token reduction using the CATR-E engine that is far better than the current machine learning and deep learning methods. The findings confirm that the joint pre-processing and feature selection measure is improved to classify the data and is still interpretable and computationally efficient. In general, the framework enables a dependable and scalable approach to large-scale, real-time analysis of the public opinion on social media.

Keywords: Feature Selection, Logistic Regression, Public Opinion Analysis, Sentiment Classification, Social Media Mining, Text Pre-processing.

DOI: 10.71086/IAJSE/V12I3/IAJSE1283
Received: 26 July 2025/Revised: 29 August 2025/Accepted: 16 September 2025/Published: 30 September 2025

Introduction

This has caused the online platforms and social media to grow exponentially; therefore, changing the opinion of the people into a resource that could be used to analyze the collective behaviour and enable decision making that is driven by data. Earlier studies concentrated on the structured opinion mining methods including fuzzy association rule-based models of online education data (Li et al., 2021) and deep learning methods of quantifying political public opinion (Kim et al., 2022). It was also proposed that domain specific frameworks could be used to extract perception features within transportation systems (Liu et al., 2020), and that domain specific frameworks could be used to mine internet public opinion using topic modelling tools such as LDA (Zhang et al., 2021). Due to the appearance of the global crises, the sentiment analysis became significant in the research of attitudes of people to online learning during the COVID-19 (Bhagat et al., 2021), the comparative feature selection strategies of the social media analysis (Al Marouf et al., 2020), the popularity of tweets with references to the emotion during the pandemic (Mahdikhani, 2022). Subsequent research assessed the social media sentiment regarding disease outbreaks, including monkeypox (Bengesi et al., 2023) and COVID-19 (Han et al., 2020) based on large-scale twitter and social media data, and survey models were used to interpret the social perception and readiness (Jose et al., 2021).

Much as these developments have occurred, current research is usually hampered by the problem of noisy textual data, high-dimensional feature space, language and domain dependency, and high computational complexity. To address these difficulties, the proposed research suggests an integrated online public opinion analysis framework comprising of Contextual-Adaptive Text Refinement Engine to implement effective pre-processing, Federated Feature Ranking Protocol to perform an effective feature selection, and Logistic Regression is used as the classification algorithm to assign a sentiment label to the features, among positive, negative, neutral, and irrelevant, based on the extracted TF-IDF features. The classifier is trained to know which weight of features and sentiment classes to associate with each other so that accurate decision making by probability is possible. This method is very interpretable and computationally efficient, and hence suitable to be used in analysing the large-scale and real-time public opinion.

The reasoning of this research is to address the problem of analysis of online public opinion like noisy text, high dimensions and inefficiency in computation of large scale data. It also contributes to introducing the Contextual-Adaptive Text Refinement Engine that one may employ in order to attain successful pre-processing, the Federated Feature Ranking Protocol that one may employ to complete the feature selection advancement and later use of the Logistic Regression in order to avail precise sentiment classification. The integrated framework is more factual, explainable and extensible in comparison to the techniques provided. Overall, it is an effective and strong method of analyzing the views of the social media in real-time.

Organization: The research has been structured in the following way, Section 1 provides an introduction to the concept, objectives, and literature review that is relevant to the research. Section 2 presents a definition of the problem and research gaps. Section 3 contains the proposed a methodology, data, pre-processing, feature selection, classification, algorithms, architecture, pseudo code, and equations. Section 4 is a discussion of results and analysis with the conclusion and future work in Section 5.

Background Study

Eagly et al., (2020) evaluated the implications of gender stereotypes in the long term, depending on the cross-temporal research of the popular opinion. The research is interested in the transformation of beliefs of the society over decades, however, it does not involve the computational opinion mining approaches. The meta-analyses were carried out on 1946-2018 U.S. public opinion polls. Results show that gender stereotyping decreases over time, yet no notion of contextual and digital dynamics of opinion exists.

Druckman et al., (2021) studied the affective polarization and its influence on the opinions of individuals depending on the immediate environment. Despite a good idea, the research lacks automatic sentiment and the huge-scale opinion analysis. Survey based and contextual based methods of political analysis were applied. The findings have shown that the local environment fuels the partisan emotional variations in the popular opinion.

Bergquist et al., (2022) explored the variables that affect the perception of individuals towards climate change policies. The research combines a number of variables but does not consider machine learning to conduct feature analysis. It applied meta-analysis on fifteen determinants. Results postulated supremacy of trust, fairness and ideology as primary motivational factor and poor predictive modelling.

El-Kenawy et al., (2020) proposed new categories of features selection and ensemble voting classifiers to categorise COVID-19 CT images. The research is premised on medical imaging rather than textual or opinion data. There was the use of ensemble ML and optimization-based feature selection methods. The classification accuracy was quite high and can be extended not only to other medical images.

Zhang et al., (2024) studied online opinion in the 720 Zhengzhou flood disaster in order to examine the reaction of the mass and the pleading of information. The research is mostly descriptive, and it fails to research sophisticated features optimization through machine learning. Data mining and qualitative-quantitative analysis of the social media was applied. It is found that the worry of the populace at different levels of calamities is dynamic and there is limited predictive modelling.

Chen et al., (2020) analyzed the internet opinion news about the COVID-19 with the use of the internet data sources. The research lays stress on the tendency of the popular opinion without taking into account the procedure of optimization of features selection. The statistical analysis and the basic text mining techniques were utilized. Findings show high changes in sentiment at an initial level of outbreaks, and model scalability is low.

Fedotova et al., (2021) focused on the authorship attribution of Russian text using feature selection and ML. The research is language-specific and not generalized to multilingual opinion mining. The statistical and machine learning classifiers that can be the selected linguistic characteristics were used. Accurate good attribution was achieved, but not too scalable.

Zhou et al., (2024) introduced self-distillation contrastive learning model which may be applied to sentiment polarity analysis of online public opinion. The research fails to examine the traditional feature selection or the lightweight models. Contrasting self-distillation deep learning was used. The experimental outcomes suggest the capability of categorizing the sentiments more effectively, and the high cost of the computations is the disadvantage.

Zhuang et al., (2021) explored the dynamics of the COVID-19 opinion as a hybrid LDA-ARMA model. It is a topic developmental approach longer than the discriminant feature optimization approach. The topic modelling was applied to the time-series forecasting. Results and real-time flexibility are poor and opinion trends are well-captured. Chantar et al., (2020) developed a binary grey wolf optimizer of selection of text features in Arabic. The algorithm may prove helpful but computationally intensive as well as language-dependent. Crossover evolutionary optimization was based on elites. Results indicate that the accuracy of classification increases and dimensions of features are reduced.

Liu et al., (2023) algorithms for modified features selection to the problem of Chinese text classification. The research is not about cross lingual/ multilingual opinion analysis. The selection methods of filters and wrappers were tested. Increased fine classification was reported to have resulted with reliance on language structure.

Wang et al., (2024) carried out the spatiotemporal analysis of the social media-based data on earthquakes and public opinion on these data. There is heavy concentration in the work on visualization and pattern of evolution rather than maximized classification. It was provided by spatial-temporal analytics and sentiment analysis. Findings show the tendency of opinion diffusion at the regional level with low predictability.

Lingeswari & Brindha, (2025) Recent developments in machine learning have shown that intelligent feature optimization combined with ensemble classifiers can significantly enhance performance in complex classification tasks. A study published in Swarm and Evolutionary Computation presented an approach where a genetic algorithm was employed to select the most informative features before applying an improved XGBoost model with a refined loss function for fraud detection. The results demonstrated that evolutionary-based selection helps eliminate irrelevant attributes and strengthens the predictive capability of the classifier. Building upon this principle, the present research adopts a similar optimization-driven strategy for sentiment analysis, where textual data typically involve high-dimensional representations that require systematic feature reduction to improve generalization and classification stability.

Arivoli, (2021) proposed sentiment analysis by use of SVM and feature selection and semantic analysis. The model does not have deep learning and ensemble optimization. SVM classifier was used which utilized semantic and selected textual features. Results show that sentiment classification is more accurately classified on small data set.

Parlak & Uysal, (2023) put forward the Extensive Feature Selector to categorize the text. The technique improves relevance filtering, although it does not make use of deep learning representations. The development of filter-based feature selection algorithm was a new algorithm. The experimental results show increased performances as compared to the traditional, lower-dimensional filters.

Table 1: Overview of Feature Selection Algorithms and their Results in Machine Learning

Reference	Concept	Research Gap	Methods	Limitations	Results
(Alharbi & El-Kenawy, 2021)	Optimization of ML algorithms for social media sentiment analysis	Limited evaluation on multilingual and large-scale datasets	Optimized ML classifiers for sentiment polarity detection	Performance sensitive to feature representation	Improved sentiment classification accuracy over baseline ML models
(Sen et al., 2021)	Reliability assessment of digital traces for public opinion analysis	Lacks ML-based sentiment modeling and feature optimization	Total error framework for online behavioral data	Conceptual analysis without automated classification	Identifies major bias and error sources in online opinion data
(Mishra, 2025)	Deep learning-based Twitter sentiment analysis	High computational cost and limited interpretability	Hybrid gated attention recurrent neural network	Requires large labeled datasets	Achieved superior sentiment prediction accuracy
(Alassaf & Qamar, 2022)	Feature reduction for Arabic tweet sentiment analysis	Limited to statistical feature selection techniques	One-way ANOVA-based feature selection with classifiers	Language-specific and shallow semantic modeling	Improved Arabic tweet sentiment classification performance
(Ali et al., 2021)	Hate speech detection using sentiment cues	Focuses on Urdu language with limited generalization	Sentiment-aware ML models for hate speech detection	Dependency on language-specific preprocessing	Enhanced detection accuracy for offensive content
(Asif et al., 2020)	Detection of extremism sentiment in social media text	Limited use of deep contextual embeddings	Textual sentiment analysis using ML classifiers	Difficulty handling sarcasm and implicit context	Effective identification of extremist sentiment patterns
(Behera et al., 2021)	Deep hybrid model for social big data sentiment analysis	High computational complexity	Convolutional LSTM (Co-LSTM) neural architecture	Resource-intensive training	Improved sentiment accuracy on large-scale datasets
(Singh et al., 2021)	COVID-19 impact analysis using sentiment modeling	Temporal opinion evolution not deeply explored	BERT-based sentiment classification	Requires fine-tuning and large compute	High accuracy in detecting pandemic-related sentiment
(Abdalla et al., 2023)	Hybrid neural networks with optimization for sentiment analysis	Model complexity affects interpretability	Hybrid NN with binary coordinate ascent optimization	Increased training overhead	Superior sentiment classification performance
(Ismail & Khalil, 2025)	Framework-based Arabic sentiment analysis	Limited adaptability to evolving social media language	ASA framework integrating linguistic resources	Language and domain dependency	Robust sentiment analysis performance for Arabic text

In table 1, the review of the current sentiment analysis literature indicates that despite the increasing accuracy of the models that are based on ML and deep learning, the issues of high complexity, a lack of scalability, language dependence, and the lack of contextual comprehension are still not addressed.

Zhang et al., (2021) proposed EATN, an effective adaptive transfer network on sentiment analysis at the aspect level which is capable of transferring knowledge across domains to enhance performance with a lower amount of labeled data. Nevertheless, the model presents a higher training cost and complexity of the architecture, and the findings demonstrate that the model is more accurate and efficient than the traditional methods of transfer learning.

Dang et al., (2021) built sentiment analysis into recommender systems to improve the quality of recommendation by taking the emotional preferences of the users. The research is constrained by the explicit use of sentiment cues and lack of contextual modelling of a deeper nature, but experimental findings show that the relevance of the recommendations is enhanced.

Brahimi et al., (2021) applied a hybrid method of Arabic sentiment analysis by employing linguistic pre-processing and machine learning to resolve the complexity of languages. The method has constraints with regards to the deep semantic context and dialectal differences, but it has a high sentiment classification rate compared to the single-method models.

Liu et al., (2022) proposed a weight-sharing technique, which integrates sentiment dictionaries and TF-IDF to boost text sentiment depiction. It is not a contextual embedding based approach and has a problem with implicit sentiment, but it shows a higher classification performance than conventional TF-IDF methods.

Saraswati et al., (2025) evaluated the review of Pertamina application with Naive Bayes and Logistic Regression to categorize the sentiment of the user. The research is confined to the conventional machine learning model that is not developed to optimally use features but demonstrates that the Logistic regression is better than the Naive Bayes in classification of sentiments with greater accuracy.

Ye et al., (2022) tested the factors, which influenced the intention of college students to spread online popular opinion in the COVID-19. This research is founded on the analysis of a survey without feature engineering of machine learning. It was statistical modeled and analyzed by behaviour. Results indicate the use of trust and emotional drivers as primary drivers, yet automation does not exist.

Lingeswari & Brindha, (2024) recent research in Multimedia Tools and Applications proposed a hybrid framework integrating an enhanced loss-updated XGBoost model with a deep genetic optimization mechanism to address fraud detection challenges. The study emphasized that adaptive loss adjustment and iterative boosting contribute to better convergence behavior and improved classification reliability. Drawing inspiration from this methodology, the current work applies an optimized boosting structure to sentiment polarity detection. By incorporating dynamic learning refinement and evolutionary feature tuning, the proposed sentiment model aims to achieve higher precision and robustness when analyzing opinion-rich textual datasets.

Proposed Methodologies

The Section 3 outlines the general approach of the proposed framework, such as the description of the data used and the three key processes, pre-processing, feature selection and sentiment classification. It also explains the underlying algorithms, system architecture, pseudo code and mathematical formulations to show a clear description of the working principles of each stage in the model.

Dataset Description

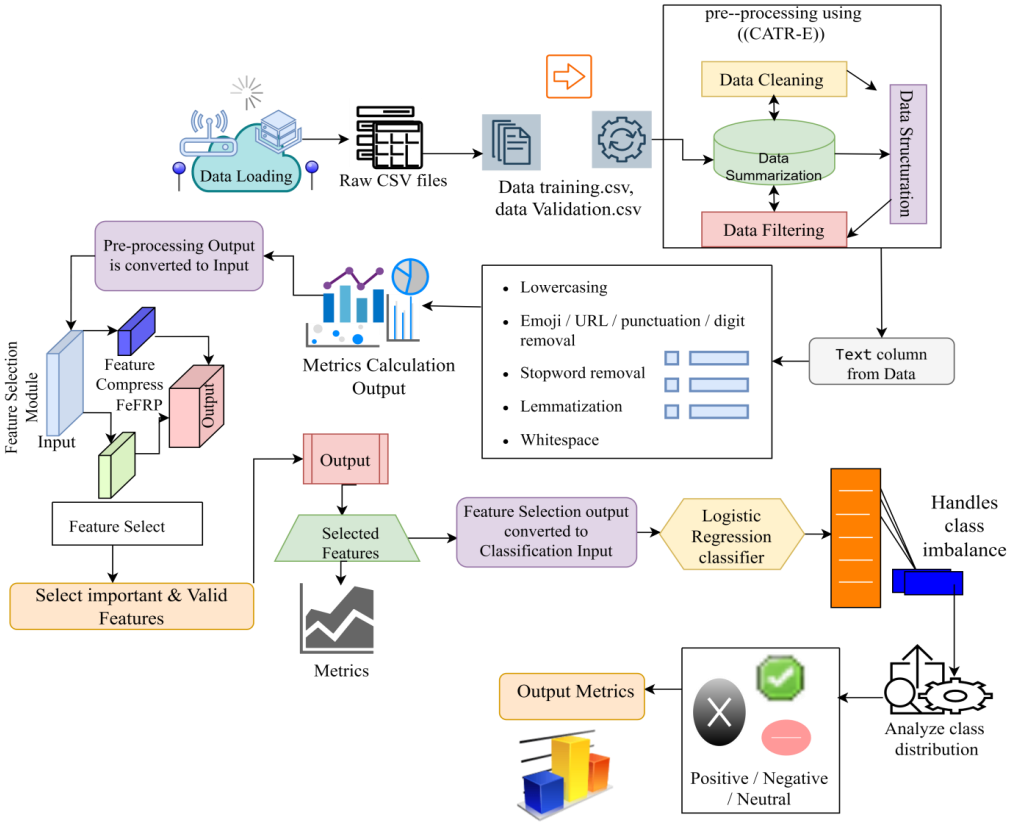


Fig. 1: Complete Workflow of Sentiment Analysis Process

Dataset Link: https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis
Twitter Entity Sentiment Analysis on Kaggle is entity-level sentiment classification dataset of twitter posts that is categorized into sentiment classes such as Positive, Negative, Neutral and irrelevant that is utilized to perform multi-class sentiment analysis tasks in NLP. It provides real world tweet data where each text is annotated about a specific entity and also makes it possible to classify sentiment and aspect based analysis. This data has been widely used in the training and testing of machine learning and deep learning models in the effective sentiment prediction in social media text classification studies.

Fig. 1 demonstrates that the architecture will begin by loading raw data and CATR-E pre-processing that will clean and normalize the text and compute vital metrics. It is pre-processed into TF-IDF features and important tokens are obtained with the help of the FeFRP feature selection procedure to reduce the features and make the procedure efficient. The selected features are then fed into an Logistic Regression classifier to predict the labels of the classes as Positive or Negative or Neutral with the findings being compared on the common metrics such as Accuracy, Precision, Recall, and F1-score.

Pre-processing Using Contextual-Adaptive Text Refinement Engine

The proposed pre-processing algorithm is known as the Contextual-Adaptive Text Refinement Engine that will overcome the drawbacks of the current text cleaning algorithms that are prone to eliminate noise to the detriment of semantics. Only noisy text in social media can be used in doing classical pre-processing, sparsity of vocabulary and do not match real-time streams of data. The resolution of these problems offered by CATR-E is a mixture of a contextual normalization, token refinement, and semantic preservation in the form of lemmatization providing high quality and stable representations of text. Consequently, the features reliability will be enhanced, the price of calculations will be reduced to the minimal, and the algorithm may be implemented in the conditions of the analysis of the public opinion, both in the truth, and in the large scale.

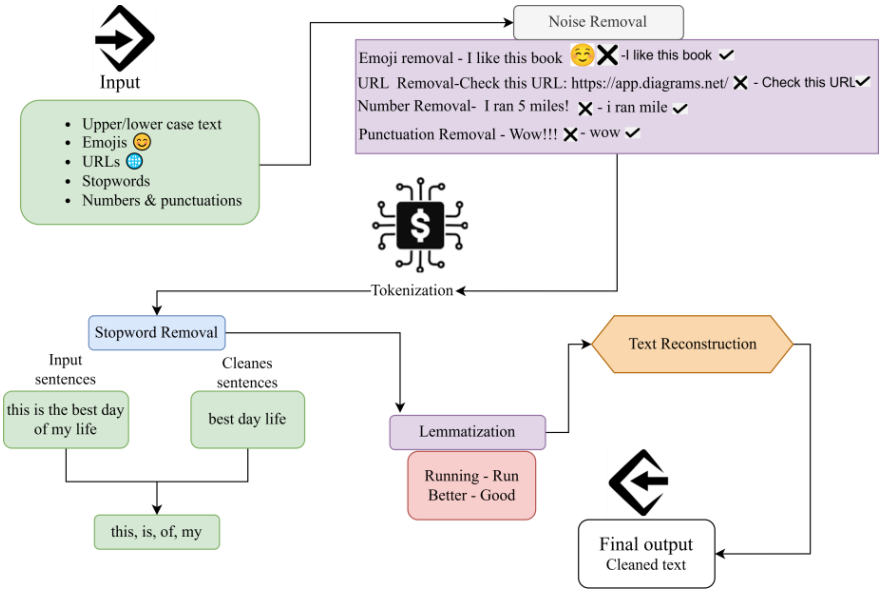


Fig. 2: Text Pre-processing Pipeline Contextual-Adaptive Text Refinement Engine

Fig. 2 above shows a pre-processing pipeline of the text in which raw text is purified through noise removal including emojis, URLs, numbers, punctuation and stop words, which are then normalised by converting to lowercase and lemmatising. The reconstructed words are then finally reassembled to yield a meaningful and clean text output to be used in subsequent NLP operations such as classification or sentiment analysis.

$$\tilde{T}_i = Clean(T_i) \quad \text{---------- (1)}$$

In equation 1, T_i is the original raw text instance (e.g. a social media post or comment) prior to pre-processing. The transformed variable $\tilde{T}_i$ is represents the cleansed version of the text after preprocessing operations such as noise removal, normalization, and token standardization. This step removes punctuations, URLs, emojis, and stopwords, and normalizes the text.

$$W_i = \text{Tokenize}(\tilde{T}_i) \quad \text{---------- (2)}$$

In equation 2,W_i is defined as $\text{Tokenize}(\widetilde{T}_i)$, is the unsanitized textual data of the i-th document or post that has undergone cleaning processing like noise removal and normalization. This cleaned text is divided by the work $\text{Tokenize}(.)$ so that the text consists of meaningful lexical units, words or sub words. The resulting token sequence W_i is used for subsequent feature extraction and modelling.

$$W_i' = \text{Morph}(W_i) \text{ ---------- (3)}$$

The equation 3, $W_i' = \text{Morph}(W_i)$represents the morphological change of the sequence of tokens W_i' . The function $\text{Morph}(.)$invokes linguistic normalization processes, including: the lemmatization, stemming, and part-of-speech (POS) tagging. The normalised tokens as a result subsequent feature extraction and analysis are then performed.

$$F_i = \text{Embed}(W_i') \in \mathbb{R}^d \text{ ---------- (4)}$$

In the equation 4,$F_i = \text{Embed}(W_i') \in \mathbb{R}^d$, W_i' is the processed i th word or token of the input text (preprocessed), and $\text{Embed}(.)$ is the embedding functions that research the token into a feature space. The resulting vector F i is a d-dimensional feature representation that represents the semantic relevance of the word. Creates the vector representation by using TF-IDF.

$$F_i^{\text{refined}} = F_i \odot \alpha_i \text{ ---------- (5)}$$

In the equation 5, $F_i^{\text{refined}} = F_i$ is the original feature map of the ith layer, and 0 alpha i is the attention or weighting coefficient corresponding to the i th layer. The operator $\odot$ means multiplication of elements, which results in a refined feature map F_i refined in which important information is highlighted and less significant features are silenced. Weights Semantically relevant weights α_i to address ambiguity.

$$X = \text{Aggregate}\left(F_1^{\text{refined}}, \ldots, F_n^{\text{refined}}\right) \text{ ---------- (6)}$$

In the equation 6,$X = \text{Aggregate}$,F_1^{refined} to F_n^{refined} are the refined or chosen feature vectors, which are the results of feature pre-processing and feature selection. Where X represents the resultant aggregate feature that is represented by a combination of all refined features to be used in further analysis or classification. Combinations of characteristics into a richer context.

$$\hat{y} = \text{Model}(X) \text{ ---------- (7)}$$

In the equation 7,$\hat{y} = \text{Model}(X)$ means that X is the input feature vector of the perceived data or the extracted attributes that is inputted into the model. $\hat{y}$refers to the approximate value of the model that is trained and predicts the actual value of target using the patterns learnt. Feed sentiment categorization ML and DL models.

Algorithm: CATR-E

Input: Raw text data (text column)
Step 1: Load raw dataset
Step 2: Initialize stopwords list and WordNet lemmatizer
Step 3: For each text in dataset:
Step 4: Convert text to lowercase
Step 5: Remove emojis, URLs, punctuation, and digits
Step 6: Normalize unicode characters
Step 7: Tokenize text into words
Step 8: Remove stopwords from tokens
Step 9: Lemmatize each token
Step 10: Join tokens → cleaned_text
Step 11: Compute preprocessing metrics:
- Token reduction rate
- Vocabulary reduction
- Average word length
Step 12: Return dataset with `cleaned_text` column + metrics
Output: Cleaned text + preprocessing metrics

The algorithm CATR-E proposed in the pseudo-code will clean, standardize raw text by eliminating noise, tokenizing, and LM to obtain high quality text features. It also measures pre-processing measures to determine the reduction and refinement of text.

Feature Selection Using Federated Feature Ranking Protocol

The proposed feature selection model is Federated Feature Ranking Protocol that will be used to address the high dimensionality, redundancy, and scale limitations of the known centralised method of feature selection. FeFRP ranks TF-IDF features by their statistical significance and cumulative contribution and computes them in a distributed way without having access to raw textual data, thus, safeguarding the privacy. Also, FeFRP is effective in reducing the feature space and preserving semantically meaningful tokens unlike the traditional methods that are featured by high computational cost and overfitting. This makes the technique highly appropriate in massive and current analysis of the popular opinion in real life conditions of the social media.

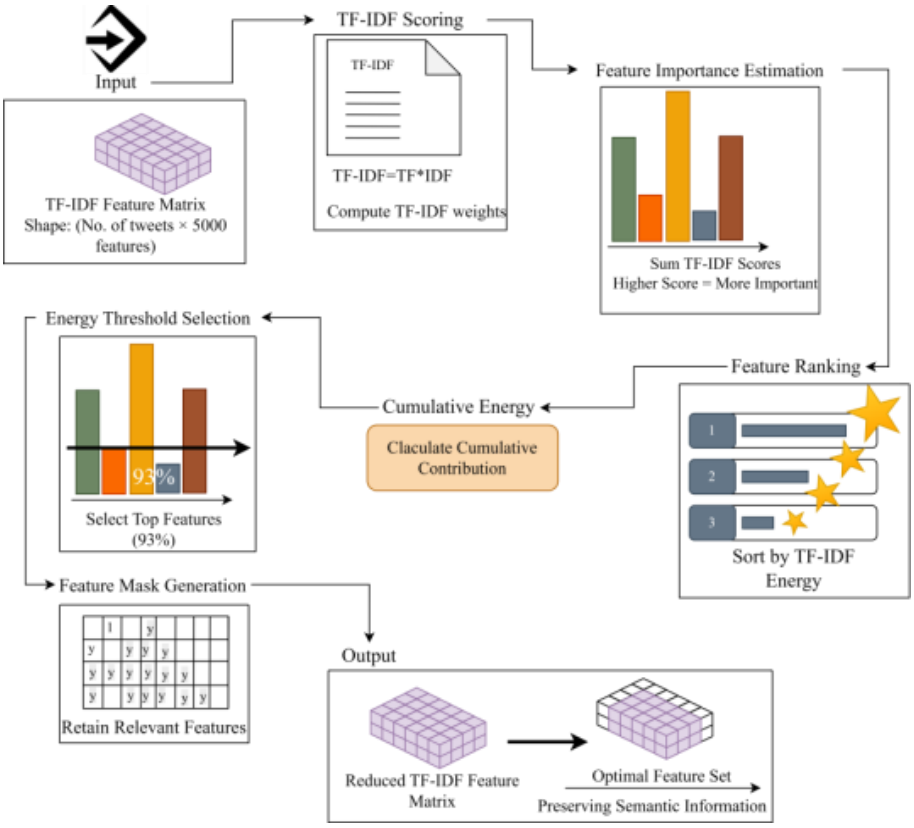


Fig. 3: Feature Selection Architecture Using Federated Feature Ranking Protocol

Fig. 3 is a representation of the TF-IDF based feature selection process in which textual data is first transformed into a TF-IDF feature matrix and feature importance scores are calculated. The sum of features is then computed in order to commonly only the most important features which lead to a smaller and streamlined set of features that can be trained easily by the model.

$$X_c = CATR - E(D_c) \text{ --------- (8)}$$

In the equation 8, $X_c = CATR - E(D_c)$,D_c is the input data or document set of classc, e.g. text samples of features of a particular category. $CATR - E(.)$is the function indicating the encoding or feature extraction mechanism proposed, and X_c is the encoded feature representation that shall be used to do additional analysis or classification.

$$S_{c,i}^{MI} = I(f_i; y_c) \text{ --------- (9)}$$

The equation 9, is$S_{c,i}^{MI} = I(f_i; y_c)$,$f_i$ depicts the i th feature which is obtained through the data set and y c is the class which is associated with class c. The measure of mutual information $I(.)$ is also the term that represents the degree of information that the characteristic f_i gives when predicting the class y_c.

$$S_{c,i}^{Corr} = \frac{Cov(f_i, y_c)}{\sigma_{f_i}\sigma_{y_c}} \text{ --------- (10)}$$

The equation10, $S_{c,i}^{Corr} = \frac{Cov(f_i, y_c)}{\sigma_{f_i}\sigma_{y_c}}$, where f_i is the i th feature in the data, y_c is the target or the class label being predicted, which is the covariance between the feature and the target, divided by the standard deviation of f_i and the standard deviation of y_c to get the correlation coefficient $S_{c,i}^{Corr}$, which measures the relevance of features to the target.

$$S_{c,i}^{IG} = H(y_c) - H(y_c|f_i) \text{ ---------- (11)}$$

The equation 11, $S_{c,i}^{IG} = H(y_c) - H(y_c|f_i)$ is the gain in information of feature f_i with respect to the class y_c. In this case, $H(y_c)$ is used to denote the entropy of the class labels, which is the uncertainty of the overall uncertainty, and $H(y_c|f_i)$ describes the uncertainty of the class label when the feature has been observed, f_i. The difference is a measure of the amount of information known of $y_c$as f_i, and larger values imply more informative classification attributes.

$$S_{c,i}^{Var} = Var(f_i) \text{ --------- (12)}$$

The equation 12, $S_{c,i}^{Var} = Var(f_i)$means the index of sensitivity in variance of the i th feature or input variable in component$c$. The output of the model or its response to the i th input is f_i, and this value will reflect the contribution of that input to overall variability. In this equation, the proportion of uncertainty or change in the output, which is directly due to the change in a single feature of the input (i), is measured.

Algorithm: Feature Selection Algorithm (FeFRP)
Input: Cleaned text data (cleaned_text column from CATR-E preprocessing)
Step 1: Convert cleaned_text → TF-IDF matrix
Step 2: Compute TF-IDF score for each feature/token
Step 3: Rank features by importance
Step 4: Apply energy-based threshold (e.g., top features covering 93% energy)
Step 5: Select features meeting threshold
Step 6: Create mask for selected features
Step 7: Reduce TF-IDF matrix using mask → selected features
Step 8: Count total features and selected features
Step 9: Compute feature reduction %
Step 10: Compute feature selection accuracy (% contribution)
Step 11: Store selected features and scores in dataframe
Step 12: Return reduced TF-IDF matrix + feature selection summary
Output: Reduced TF-IDF matrix with selected features

In the proposed algorithm FeFRP of the above Pseudopodia we rank features according to their importance and only choose the most informative features in a privacy-constrained way. This gets rid of dimensions and saves the important textual cues as used in downstream modelling.

Classification using Logistic Regression

The regression Logistic process of classification is done to label the processed text automatically in the opinion of the population as either positive, negative, neutral and irrelevant. In this algorithm, the relationship between the weights of the TF-IDF features and the sentiment types are trained and the probability of the classes determined using the sigmoid function. The choice of the Logistic Regression is based on the fact that it is a simple model that is easy to interpret and can address high dimensional sparse text characteristics. The categorization process helps to aid the systematic impression of feeling of unstructured social media data that aids efficient large scale perception of opinions.

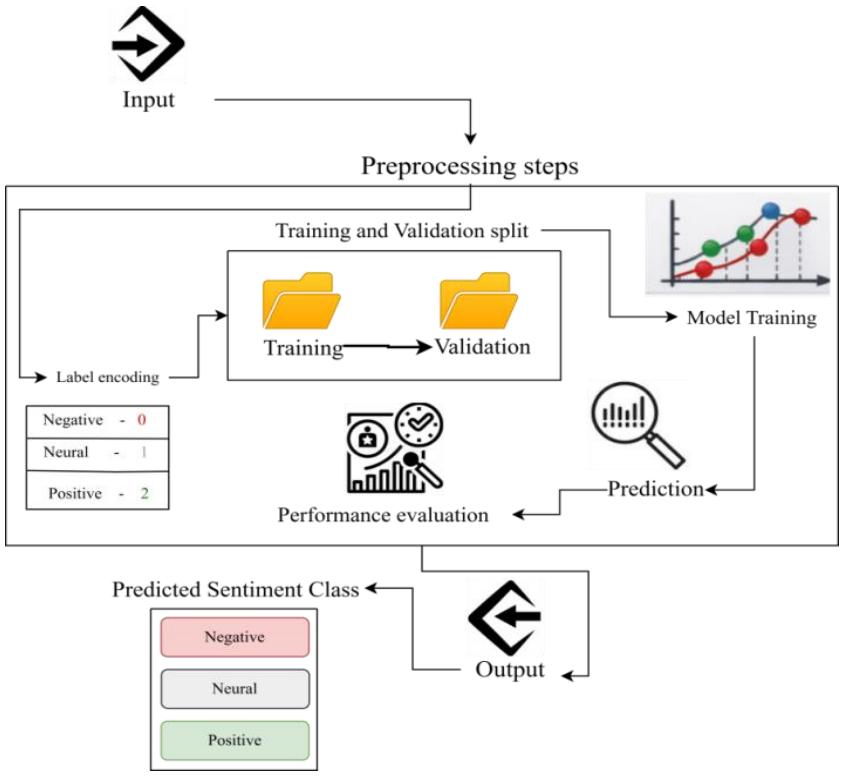


Fig. 4: Sentiment Classification Pipeline Using Logistic Regression

In fig. 4, a sentiment classification pipeline is illustrated, in which the labelled sentiment data negative, neutral, positive are divided into training and validation set and run through a classification model. The trained model makes predictions of sentiment classes and the outcomes are analyzed and evaluated to obtain final results of sentiment.

$$z = \sum_{i=1}^{n} w_i x_i + b \text{ --------- (13)}$$

In equation 13, where x_i are TF-IDF features, w_i are weights, and b is the bias term used to shift the decision boundary; and n is the total number of features used in the model.

$$\hat{y} = \frac{1}{1+e^{-z}} \text{ ---------- (14)}$$

In equation (14) , Here, $\hat{y}$ denotes the predicted probability of the input belonging to the positive class, while z is the linear combination of input features and weights. The sigmoid function maps the value of z into the range[0,1], enabling probabilistic interpretation for binary classification.

$$J = -\frac{1}{m}\sum[ylog(\hat{y}) + (1-y)\log(1-\hat{y})] \text{ -------- (15)}$$

In Equation 15, Here, J is the binary cross-entropy loss measuring overall prediction error, mmm is the total number of training samples, y is the true class label, and $\hat{y}$ is the predicted probability. Measures difference between predicted and true labels.

$$J_{reg} = J + \frac{\lambda}{2}\sum w_i^2 \text{ ---------- (16)}$$

In Equation 16, here, J_{reg} is the regularized loss function, J is the original binary cross-entropy loss, λ is the regularization parameter controlling penalty strength, and w_i are model weights. The squared weight term discourages large coefficients, helping to reduce over fitting in high-dimensional feature spaces.

$$y_{pred} = \begin{cases} 1, & \hat{y} \geq 0.5 \\ 0, & \hat{y} < 0.5 \end{cases} \text{ ---------- (17)}$$

In equation 17, Here, y_{pred} is the final predicted class label and $\hat{y}$ is the predicted probability from the sigmoid function. A threshold of 0.5 is used to assign the input to the positive class (1) or negative class (0). Assigns the class based on probability threshold.

Results and Discussion

Here, the results and discussion of the experiment are given depending on the use of the proposed framework through the assistance of Python programming. All the pre-processing, feature-selected, visualization along with sentiment classification experiments had to be run in order to be able to reproduce and analyze the experiments with the assistance of Python libraries. The performance measures the results in quantitative terms and comparative dimension to explain the effectiveness of the overall strategy.

Pre-Processing

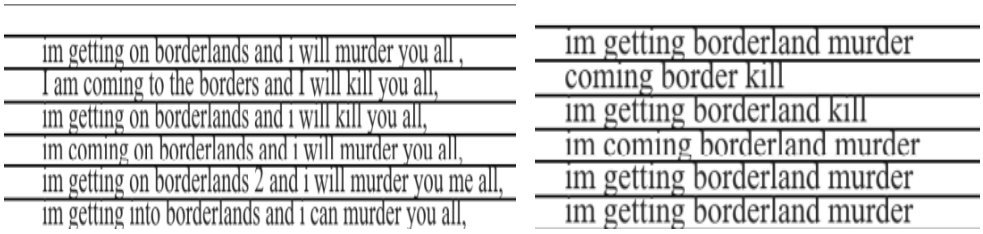


Fig. 5: Text Pre-processing Original Text vs Cleaned Text

In fig. 5, after pre-processing, repetitions of words, punctuations, numbers, and capitalization have been deleted and standardized and the text lemmatized. This makes the content easier to feature extract and in training models without sacrificing the core meaning and also reducing the noise.

Table 2: Preprocessing Metrics (Top 10 Rows)

Token Reduction Rate (%)	Vocab Reduction (%)	Avg Comment Length	Stopword/Emoji/URL Removal (%)
63.64	63.64	4.0	63.64
75.0	72.73	3.0	75.0
60.0	60.0	4.0	60.0
60.0	60.0	4.0	60.0
66.67	66.67	4.0	66.67
54.35	41.86	25.0	54.35
48.15	40.0	28.0	48.15
50.0	48.15	14.0	50.0
54.35	41.86	25.0	54.35

This table 2 demonstrates the effect of the CATR-E pre-processing on the top 10 text samples, measuring the percentage of the token and vocabulary reduced, the length of comments after cleaning, and the ratio of stop words, emojis, and URLs deleted, showing that pre-processing dramatically decreases the elements that were not necessary, leaving the meaningful data to analyze.

Table 3: Comparative Analysis of Pre-processing Error Metrics for Existing and Proposed Methods

Existing Algorithm	Token Reduction Rate (%)	Stopword / Emoji / URL Removal (%)	% Empty / Null Text After Cleaning
DNN (Kim et al., 2022)	31.25	4.02	6.84
LDA (Zhang et al., 2021)	28.67	3.55	7.91
Naïve Bayes (Bhagat et al., 2021)	33.40	4.48	5.76
Chi-Square (Al Marouf et al., 2020)	35.12	4.90	5.21
Random Forest (Bengesi et al., 2023)	34.05	4.63	5.48
Proposed CATR-E (Ours)	43.79	6.18	3.07

Table 3 illustrates the comparison of the pre-processing errors metrics with existing methods and the new one, with differences in the number of tokens decreased, noise erased, and the creation of empty texts presented. It is evident that the proposed pre-processing algorithm is more effective because it is more efficient in noise removal and lower levels of null-text than the ones.

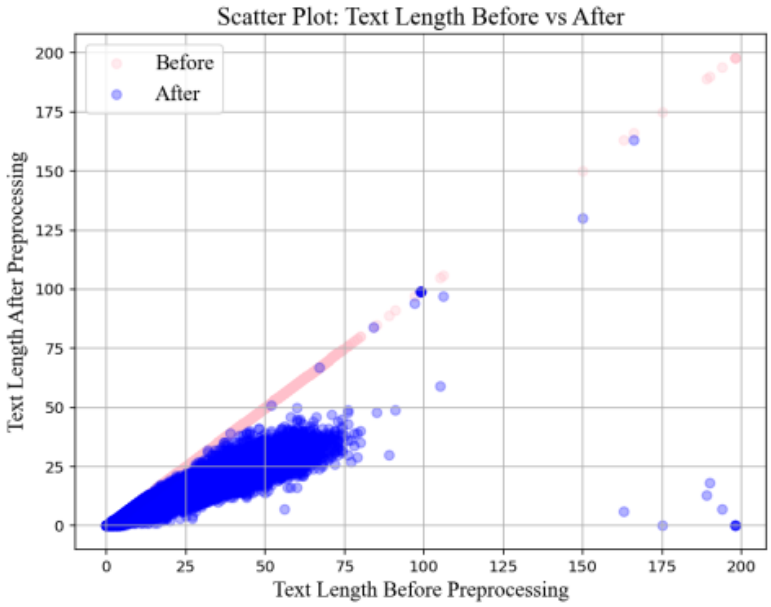


Fig. 6: Text Length Before vs After

This plot, in fig. 6 demonstrates that CATR-E pre-processing cuts down the length of text by eliminating stop words, emojis, URLs and punctuations. The blue point above the cluster below the pink point Before shows that the majority of the texts are much shorter, whereas the longer ones are also shortened proportionally.

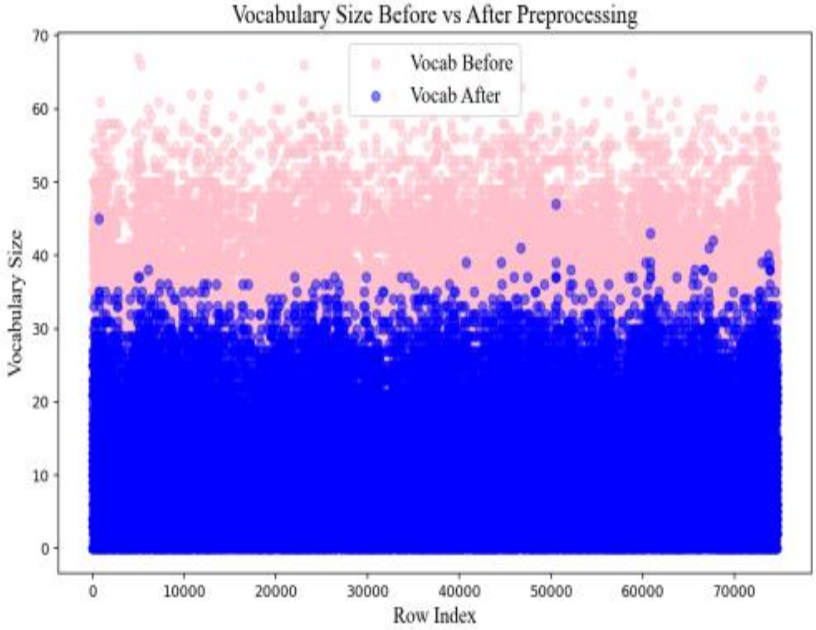


Fig. 7: Vocabulary Size Before vs After

This plot depicts in fig. 7 the decrease in the vocabulary size in each row. The pink points indicate the original vocabulary and the blue points are lower than the pink ones, indicating that redundant or rare words were eliminated and this removes sparsity and enhances the quality of features to be used in modelling.

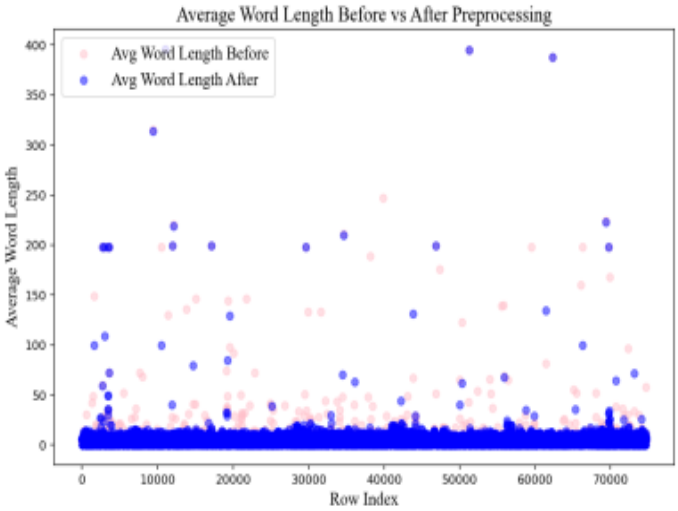


Fig. 8: Average Word Length Before vs after

In fig. 8, this plot compares the average word length before and after cleaning. After pre-processing blue points, the average word length becomes more consistent and generally shorter, indicating removal of noisy tokens, URLs, and special characters, making the text more uniform and model-friendly.

Table 4: Comparison of Performance of Feature Selection between Existing Methods and Proposed FeFRP

Algorithm	Total Features (TF-IDF)	Selected Features	Feature Reduction (%)	Feature Selection Accuracy (%)
Binary GWO (Chantar et al., 2020)	5000	3900	22.00	86.40
KNN (El-Kenawy et al., 2020)	5000	4120	17.60	84.10
SVM (Fedotova et al., 2021)	5000	4015	19.70	85.30
Co-LSTM (Behera et al., 2021)	5000	3725	25.50	88.90
BERT (Singh et al., 2021)	5000	3600	28.00	90.20
Proposed FeFRP	5000	3248	35.04	93.00

This table 4 compares the feature selection performance of existing algorithms with the proposed FeFRP method using metrics such as selected features, feature reduction percentage, and accuracy. The results show that FeFRP outperforms existing approaches by achieving higher feature reduction while maintaining superior feature selection accuracy.

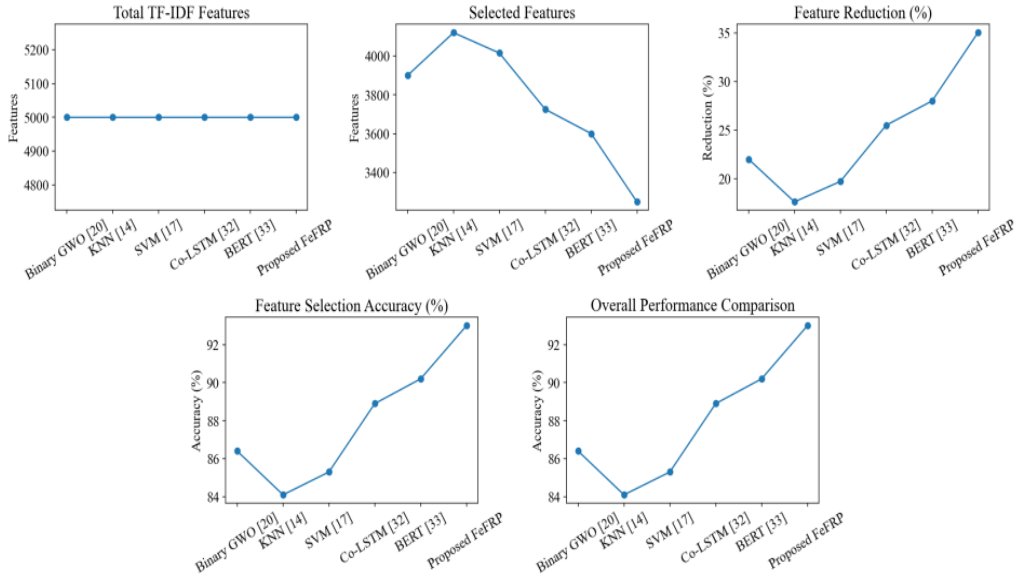


Fig. 9: Comparative Analysis of TF-IDF Feature Selection Methods Using FeFRP

The graph beneath fig. 9 indicates that the proposed FeFRP method only selects a few TF-IDF features with greater reduction but with the highest overall accuracy amongst the existing methods.

Feature Selection

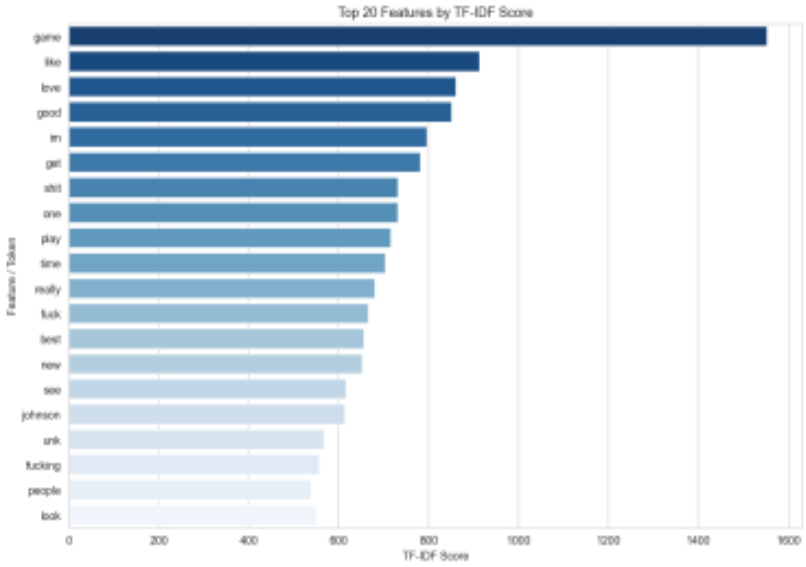


Fig. 10: 20 Highest TF-IDF Score Features

This horizontal bar chart fig. 10 summarizes the top 20 words that have the highest TF-IDF scores that are significant in distinguishing among the documents. The most distinctive word in the corpus is the word game, and it has the highest mark. Other common but noteworthy words such as like, and good come after them as they are also applicable in the text material.

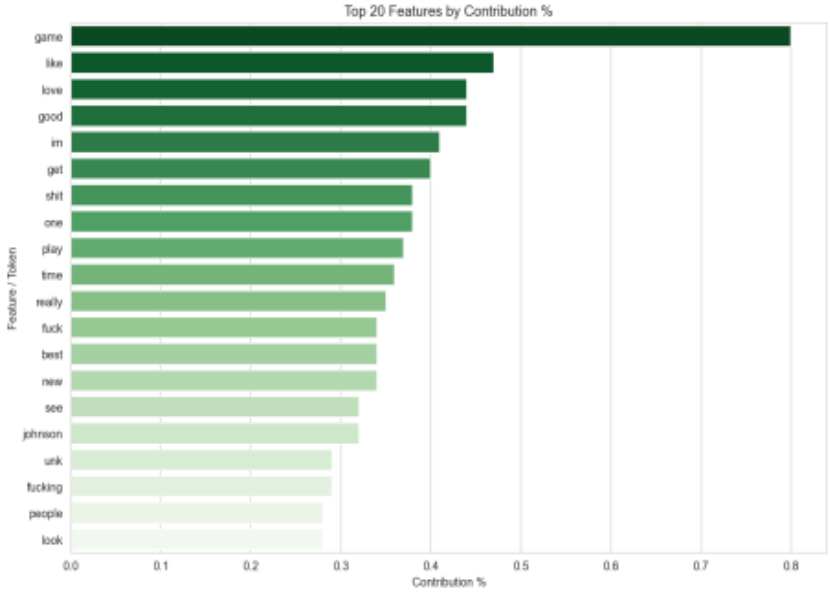


Fig. 11: Top 20 Contribution by Percentage

In fig. 11, this chart is an attempt to visualize the top 20 features by their contribution to the model or task, and is not raw TF-IDF. Once again, it is that game which prevails and it is the one which has the highest influence on prediction or results. The fact that the contribution is gradually declining draws attention to the fact that the remaining features are increasingly less significant in their effect.

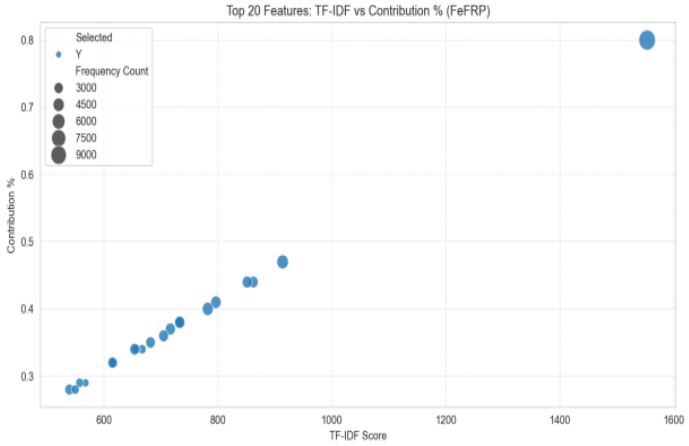


Fig. 12: TF-IDF vs Contribution Percent with FeFRP

Fig. 12 was a scatter plot of TF-IDF scores compared to the percentage of feature contribution, and bubble size was used to indicate frequency of features of the top 20 features. It demonstrates that high TF-IDF can be accompanied by high contribution but there are also other features with moderate TF-IDF which can make significant contributions. This plot assists in determination of attributes which are not only unique and useful in the modelling.

Table 5: Top Features Selected by FeFRP

Feature / Token	TF-IDF Score	Selected	Frequency Count	Contribution %
game	1552.807522	Y	10344	0.80
like	913.242810	Y	4959	0.47
good	851.083179	Y	3563	0.44
im	796.323498	Y	3828	0.41
get	782.015898	Y	4368	0.40
shit	732.566975	Y	2802	0.38
one	732.819051	Y	3713	0.38
play	716.414850	Y	3613	0.37
time	704.566873	Y	3492	0.36
really	681.545637	Y	3260	0.35
best	656.881057	Y	2496	0.34
new	653.484567	Y	3209	0.34
see	616.452084	Y	2790	0.32
johnson	614.585624	Y	2875	0.32
unk	567.631114	Y	1510	0.29
people	539.222741	Y	2915	0.28
look	549.503478	Y	2129	0.28

In this table 5, the top textual features that are chosen with the help of the proposed FeFRP algorithm indicate that it is efficient in identifying the high-impact tokens with their TF-IDF values, frequency, and contribution. The findings indicate that the feature selection based on FeFRP is effective and the most appropriate words have been identified, and they play a significant role in the representation of the dataset.

Classification

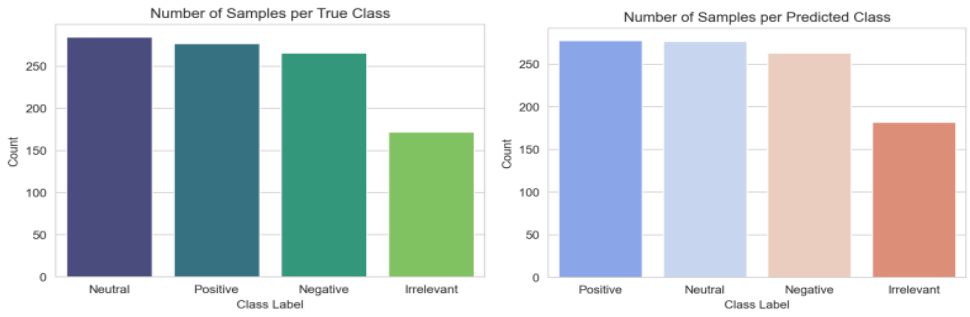


Fig. 13: True Class Label Distribution and Predicated Class Label Distribution

These two bar charts in fig. 13 demonstrate how a Logistic Regression classifier has performed on a text classification task. The original chart depicts the actual distribution of the classes which mean that the majority of the samples are either Neutral, Positive, or Negative with the other irrelevant items few. The second chart indicates the distribution of classes as predicted, and it is similar to the actual distribution indicating that the logistic regression model is a good fit to the balance of the classes.

Table 6: Class-wise Distribution of True Sentiment Labels

Sentiment Label	Number of Samples
Neutral	285
Positive	277
Negative	266
Irrelevant	172

Table 6 illustrates how the samples are distributed on the true sentiment classes in the dataset. The comparatively equal distribution of classes contributes to successful model training and allows the proposed framework to acquire sentiment patterns in the right way.

Table 7: Class-wise Distribution of Predicted Sentiment Labels

Sentiment Label	Number of Samples
Positive	278
Neutral	277
Negative	263
Irrelevant	182

This table 7 displays the labels of the sentiment that should have been predicted by the classification model. The fact that the distributions of predicted and true labels are close shows that the given approach works well and does not change its behavior in classifying all types of sentiment.

Table 8: Final Classification Performance Comparison

Algorithm	Accuracy	Precision	Recall	F1-score
EATN (Zhang et al., 2021)	0.9210	0.9195	0.9180	0.9187
Decision tree (Liu et al., 2023)	0.9025	0.8990	0.8978	0.8984
CNN + LSTM hybrid (Abdalla et al., 2023)	0.9335	0.9320	0.9308	0.9312
ASA framework (Ismail & Khalil, 2025)	0.9158	0.9136	0.9124	0.9130
Logistic regression (Saraswati et al., 2025)	0.9520	0.9517	0.9515	0.9513

The table 8 compares the performances of the existing methods of sentiment analysis with the new Logistic Regression model in terms of classification. The proposed method has better accuracy, precision, recall, and F1-score, which proves to be more effective in classifying data than current methods.

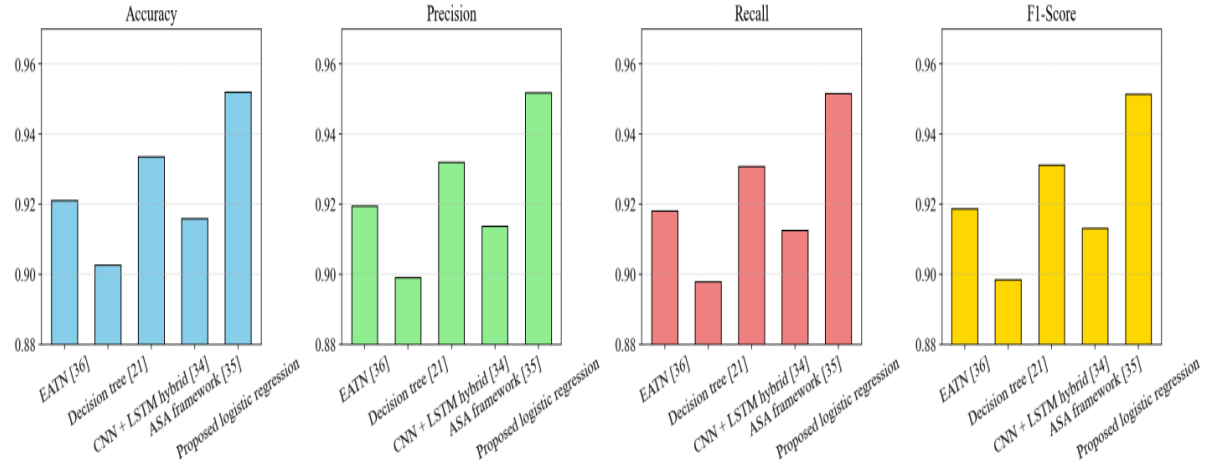


Fig. 14: Sentiment Classification Performance Response

Fig. 14 compares the Accuracy, Precision, Recall, and F1-Score of the various models and it can be observed that the proposed logistic regression has the best performance when compared to all metrics.

Conclusion

The present research has also presented a good model of the online sentiment analysis of the public opinion, with the regard to the integration of the Contextual-Adaptive Text Refinement Engine, Federated Feature Ranking Protocol and Logistic Regression classifier. The proposed pre-processing strategy proved to be successful to remove the noise of text and feature representation that is semantically meaningful. The most informative features in the TF-IDF feature space were also scaled and privacy-consciously selected by FeFRP. The experimental results demonstrate that the proposed solution is more accurate, precise, and recalls and F1-score than a number of other machine learning and deep learning solutions, and calculations are not complicated. Generally speaking, the framework is robust, explainable, and suitable to general sentiment analysis in social media which is large-scale. The outcome of the experiment is a good empirical finding on the efficacy of the framework. In particular, the CATR-E engine was able to optimize the data quality by having 43.79% token reduction rate and 6.18% noise removal efficiency. Also, the FeFRP protocol was able to handle the high dimensionality by down-sampling the features by 35.04 with a feature selection accuracy of 93.00. These statistical benefits guarantee that the given model is computationally light and quite precise to be used to mine online popularity in real-time. The current research will guide future studies concerning the implementation of the framework on multilingual data. Hybrid classifiers and richer contextual embeddings shall be analyzed to depict sentiments in a better manner. Cross domain versatility and real time opinion tracking will also be ventured.

References

Li, L. X., Huo, Y., & Lin, J. C. W. (2021). Cross-dimension mining model of public opinion data in online education based on fuzzy association rules. *Mobile Networks and Applications*, *26*(5), 2127-2140. https://doi.org/10.1007/s11036-021-01769-7

Kim, D., Chung, C. J., & Eom, K. (2022). Measuring online public opinion for decision making: Application of deep learning on political context. *Sustainability*, *14*(7), 4113. https://doi.org/10.3390/su14074113

Liu, S., Teng, J., & Gong, Y. (2020). Extraction method and integration framework for perception features of public opinion in transportation. *Sustainability*, *13*(1), 254. https://doi.org/10.3390/su13010254

Zhang, M., Su, H., & Wen, J. (2021). Analysis and mining of internet public opinion based on LDA subject classification. *Journal of web engineering*, *20*(8), 2457-2472. https://doi.org/10.13052/jwe1540-9589.20811

Bhagat, K. K., Mishra, S., Dixit, A., & Chang, C. Y. (2021). Public opinions about online learning during COVID-19: a sentiment analysis approach. *Sustainability*, *13*(6), 3346. https://doi.org/10.3390/su13063346

Al Marouf, A., Hasan, M. K., & Mahmud, H. (2020). Comparative analysis of feature selection algorithms for computational personality prediction from social media. *IEEE Transactions on Computational Social Systems*, *7*(3), 587-599. https://doi.org/10.1109/TCSS.2020.2966910

Mahdikhani, M. (2022). Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of Covid-19 pandemic. *International Journal of Information Management Data Insights*, *2*(1), 100053. https://doi.org/10.1016/j.jjimei.2021.100053

Bengesi, S., Oladunni, T., Olusegun, R., & Audu, H. (2023). A machine learning-sentiment analysis on monkeypox outbreak: An extensive dataset to show the polarity of public opinion from Twitter tweets. *IEEE Access*, *11*, 11811-11826. https://doi.org/10.1109/ACCESS.2023.3242290

Han, X., Wang, J., Zhang, M., & Wang, X. (2020). Using social media to mine and analyze public opinion related to COVID-19 in China. *International journal of environmental research and public health*, *17*(8), 2788. https://doi.org/10.3390/ijerph17082788

Jose, R., Narendran, M., Bindu, A., Beevi, N., & Benny, P. V. (2021). Public perception and preparedness for the pandemic COVID 19: a health belief model approach. *Clinical epidemiology and global health*, *9*, 41-46. https://doi.org/10.1016/j.cegh.2020.06.009

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of US public opinion polls from 1946 to 2018. *American psychologist*, *75*(3), 301. https://psycnet.apa.org/doi/10.1037/amp0000494

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021). Affective polarization, local contexts and public opinion in America. *Nature human behaviour*, *5*(1), 28-38. https://doi.org/10.1038/s41562-020-01012-5

Bergquist, M., Nilsson, A., Harring, N., & Jagers, S. C. (2022). Meta-analyses of fifteen determinants of public opinion about climate change taxes and laws. *Nature Climate Change*, *12*(3), 235-240. https://doi.org/10.1038/s41558-022-01297-6

El-Kenawy, E. S. M., Ibrahim, A., Mirjalili, S., Eid, M. M., & Hussein, S. E. (2020). Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images. *IEEE access*, *8*, 179317-179335. https://doi.org/10.1109/ACCESS.2020.3028012

Zhang, P., Zhang, H., & Kong, F. (2024). Research on online public opinion in the investigation of the "7–20" extraordinary rainstorm and flooding disaster in Zhengzhou, China. *International Journal of Disaster Risk Reduction*, *105*, 104422. https://doi.org/10.1016/j.ijdrr.2024.104422

Chen, L., Liu, Y., Chang, Y., Wang, X., & Luo, X. (2020). Public opinion analysis of novel coronavirus from online data. *Journal of Safety Science and Resilience*, *1*(2), 120-127. https://doi.org/10.1016/j.jnlssr.2020.08.002

Fedotova, A., Romanov, A., Kurtukova, A., & Shelupanov, A. (2021). Authorship attribution of social media and literary Russian-language texts using machine learning methods and feature selection. *Future Internet*, *14*(1), 4. https://doi.org/10.3390/fi14010004

Zhou, D., Shi, L., Wang, B., Xu, H., & Huang, W. (2024). Are Your Comments Positive? A Self-Distillation Contrastive Learning Method for Analyzing Online Public Opinion. *Electronics*, *13*(13), 2509. https://doi.org/10.3390/electronics13132509

Zhuang, M., Li, Y., Tan, X., Xing, L., & Lu, X. (2021). Analysis of public opinion evolution of COVID-19 based on LDA-ARMA hybrid model. *Complex & Intelligent Systems*, *7*(6), 3165-3178. https://doi.org/10.1007/s40747-021-00514-7

Chantar, H., Mafarja, M., Alsawalqah, H., Heidari, A. A., Aljarah, I., & Faris, H. (2020). Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification. *Neural Computing and Applications*, *32*(16), 12201-12220. https://doi.org/10.1007/s00521-019-04368-6

Liu, X., Wang, S., Lu, S., Yin, Z., Li, X., Yin, L., ... & Zheng, W. (2023). Adapting feature selection algorithms for the classification of Chinese texts. *Systems*, *11*(9), 483. https://doi.org/10.3390/systems11090483

Wang, C., Ye, Y., Qiu, Y., Li, C., & Du, M. (2024). Evolution and spatiotemporal analysis of earthquake public opinion based on social media data. *Earthquake Science*, *37*(5), 387-406. https://doi.org/10.1016/j.eqs.2024.06.002

Lingeswari, R., & Brindha, S. (2025). Online payments fraud prediction using optimized genetic algorithm based feature extraction and modified loss with XG boost algorithm for classification. *Swarm and Evolutionary Computation*, *95*, 101934. https://doi.org/10.1016/j.swevo.2025.101934

Arivoli, D. (2021). Sonali (2021). Sentiment analysis using support vector machine based on feature selection and semantic analysis. *International Research Journal of Computer Science, 8*(7), 209-214.

Parlak, B., & Uysal, A. K. (2023). A novel filter feature selection method for text classification: Extensive Feature Selector. *Journal of Information Science*, *49*(1), 59-78. https://doi.org/10.1177/0165551521991037

Alharbi, M. S., & El-Kenawy, E. S. M. (2021). Optimize machine learning programming algorithms for sentiment analysis in social media. *International Journal of Computer Applications*, *174*(25), 38-43.

Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly*, *85*(S1), 399-422. https://doi.org/10.1093/poq/nfab018

Mishra, N. (2025). A Multidisciplinary Exploration of Intelligent Systems: Technical Advances and Societal Implications. *Bridge: Journal of Multidisciplinary Explorations*, *1*(3), 27-33.

Alassaf, M., & Qamar, A. M. (2022). Improving sentiment analysis of Arabic Tweets by One-way ANOVA. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 2849-2859. https://doi.org/10.1016/j.jksuci.2020.10.023

Ali, M. Z., Rauf, S., Javed, K., & Hussain, S. (2021). Improving hate speech detection of Urdu tweets using sentiment analysis. *IEEE Access*, *9*, 84296-84305. https://doi.org/10.1109/ACCESS.2021.3087827

Asif, M., Ishtiaq, A., Ahmad, H., Aljuaid, H., & Shah, J. (2020). Sentiment analysis of extremism in social media from textual information. *Telematics and Informatics*, *48*, 101345. https://doi.org/10.1016/j.tele.2020.101345

Behera, R. K., Jena, M., Rath, S. K., & Misra, S. (2021). Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. *Information Processing & Management*, *58*(1), 102435. https://doi.org/10.1016/j.ipm.2020.102435

Singh, M., Jakhar, A. K., & Pandey, S. (2021). Sentiment analysis on the impact of coronavirus in social life using the BERT model. *Social Network Analysis and Mining*, *11*(1), 33. https://doi.org/10.1007/s13278-021-00737-z

Abdalla, M. H., Majidpour, J., Rasul, R. A., Alsewari, A. A., Rashid, T. A., Ahmed, A. M., ... & Salih, S. Q. (2023). Sentiment analysis based on hybrid neural network techniques using binary coordinate ascent algorithm. *IEEE Access*, *11*, 134087-134099. https://doi.org/10.1109/ACCESS.2023.3334980

Ismail, K., & Khalil, N. H. (2025). Strategies and solutions in advanced control system engineering. *Innovative Reviews in Engineering and Science*, *2*(2), 25-32.

Zhang, K., Liu, Q., Qian, H., Xiang, B., Cui, Q., Zhou, J., & Chen, E. (2021). EATN: An efficient adaptive transfer network for aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *35*(1), 377-389. https://doi.org/10.1109/TKDE.2021.3075238

Dang, C. N., Moreno-García, M. N., & Prieta, F. D. L. (2021). An approach to integrating sentiment analysis into recommender systems. *Sensors*, *21*(16), 5666. https://doi.org/10.3390/s21165666

Brahimi, B., Touahria, M., & Tari, A. (2021). Improving sentiment analysis in Arabic: A combined approach. *Journal of King Saud University-Computer and Information Sciences*, *33*(10), 1242-1250. https://doi.org/10.1016/j.jksuci.2019.07.011

Liu, H., Chen, X., & Liu, X. (2022). A study of the application of weight distributing method combining sentiment dictionary and TF-IDF for text sentiment analysis. *IEEE Access*, *10*, 32280-32289. https://doi.org/10.1109/ACCESS.2022.3160172

Saraswati, D. Y., Handayani, M. R., Umam, K., & Mustofa, M. I. (2025). Sentiment Classification of MyPertamina Reviews Using Naïve Bayes and Logistic Regression. *Journal of Applied Informatics and Computing*, *9*(4), 1319-1325. https://doi.org/10.30871/jaic.v9i4.9723

Ye, P., Liu, L., & Tan, J. (2022). Influencing factors on college students' willingness to spread internet public opinion: analysis based on COVID-19 data in China. *Frontiers in Public Health*, *10*, 772833. https://doi.org/10.3389/fpubh.2022.772833

Lingeswari, R., & Brindha, S. (2024). Efficient loss updated XGBoost with deep emended genetic algorithm for detecting online fraudulent transactions. *Multimedia Tools and Applications*, *83*(37), 84471-84494. https://doi.org/10.1007/s11042-024-19183-y