Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

A holistic evaluation of machine learning algorithms for text-based emotion detection

Syed Zafar Ali Shah [1], Omar Ahmed Abdulkader [2], Sadaqat Jan [3], Muhammad Arif Shah [4], Muhammad Anwar [5, *]

[1]*Department of Computer Software Engineering, University of Engineering and Technology, Peshawar, Peshawar, 25120, Pakistan*
[2]*Department of Computer Studies, Arab Open University, Riyadh, Saudi Arabia*
[3]*Department of Computer Software Engineering, University of Engineering and Technology, Mardan, 23200, Pakistan*
[4]*Department of IT and Computer Science, Pak-Austria Fachhochschule Institute of Applied Sciences and Technology, Haripur, Pakistan*
[5]*Department of Information Sciences, Division of Science and Technology, University of Education, Lahore, 54000, Pakistan*

ARTICLE INFO

Article history:
Received 20 September 2024
Received in revised form
21 May 2025
Accepted 8 June 2025

Keywords:
Emotion detection
Machine learning
Text analysis
Transformer models
Sentiment classification

ABSTRACT

The rapid growth of social media and text-based communication has intensified interest in emotion detection (ED) from text. Extracting emotional content from large-scale textual sources—such as social media posts, blogs, and news articles—is both challenging and critical for various applications. This study evaluates the effectiveness of traditional machine learning algorithms in text-based emotion detection by conducting a systematic literature review (SLR), expert-based evaluation, and multiple case studies. The SLR, based on seven major digital libraries, applied a five-phase selection process to identify the most relevant studies. Findings show that Support Vector Machine (SVM) is the most frequently used and top-performing model (78% of studies), followed by Naive Bayes (60%), with customized datasets preferred in 70% of the literature. The Ekman model with six emotion classes was the most common framework, while datasets with four to eight emotion categories yielded the highest accuracy. An Analytical Hierarchy Process (AHP) involving 82 industry experts ranked SVM highest in accuracy, robustness, and interpretability, followed by Naive Bayes and Random Forest. Case studies further confirmed the strong performance of SVM, Logistic Regression, and Naive Bayes, with ensemble models improving accuracy by 3% over the best individual classifier. Additionally, the study explores transformer-based models, finding that DeBERTa outperforms traditional approaches by better capturing emotional subtleties in text. Limitations of conventional models are discussed, and practical recommendations for future improvements are provided.

1. Introduction

It is difficult for an individual to process the constantly increasing volume of data manually, especially in textual form. Text mining is a set of methods used to gain valuable information and complex patterns from textual data. Text mining applications with straightforward aims typically have high enough accuracy, while those with intermediate or difficult goals usually do not, and emotion detection is one of those applications. Even a normal human being can sometimes have trouble reading someone's genuine emotions from a piece of literature. When it comes to a computer automatically detecting emotions, we can easily envision how complicated the issue is. To solve this problem, Machine Learning (ML) models need to be trained first on labeled emotional data.

Once trained, the ML models can be used to detect the relevant emotion in the given piece of text without any further manual work. In comparison, emotion detection from text is a more challenging task than Sentiment detection. Although these two terms are sometimes used interchangeably, their definitions when used in computer science differ (Munezero et al., 2014).

In contrast to sentiment, which is 'a view or opinion that is held or expressed,' according to the

* Corresponding Author.
Email Address: anwar.muhammad@ue.edu.pk (M. Anwar)

https://doi.org/10.21833/ijaas.2025.07.006
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-0615-3038
2313-626X/© 2025 The Authors. Published by IASE.

Oxford Dictionary, emotion is defined as “a powerful feeling derived from one’s circumstances, mood, or interactions with others.” The definitions of ‘emotion’ and ‘sentiment’ in the Cambridge Dictionary are ‘a thought, opinion, or concept based on a feeling about a situation, and ‘a manner of thinking about something, respectively. Sentiment is often considered a consequence of emotion (Halczak, 2023). ‘Happy,’ ‘Angry,’ and ‘Love’ are examples of emotion, and accompanying states like ‘positive,’ ‘negative,’ and ‘positive’ are their sentiments respectively. Sentiment analysis extracts subjective data from text to determine a person’s polarity of attitude toward another person, item, event, or task. On the other hand, according to psychological emotion theories, emotion detection focuses on determining how a person feels about a particular event, person, or item using some established emotion models.

In almost every aspect of daily life, emotion detection from text is used. For example, it can be used to create effective e-learning systems based on student motions, enhance human-computer interactions, monitor people’s mental health, and change or improve business strategies in response to customer emotion, detect public sentiment during any national, international, or political event, and identify potential criminals or terrorists by analyzing people’s emotions. People now collect and express their feelings through social media activities more regularly and readily.

Text remains the most popular form of communication on social media, despite the rising popularity of audio and video components. People can express their emotions through social media posts like Facebook status updates, Tweets, comments on one’s own or others’ posts, product evaluations, and micro blogs. Analyzing these texts and identifying emotions and semantics from their words can be difficult. The ability to accurately identify human emotion from text has long been a promising study area, and significant efforts have been made to create the ideal automated system. Through Text Mining (TM) techniques and Machine Learning (ML) algorithms, we have attempted to cover recent studies in the field of emotion recognition in text from the literature in this study. We give a list of the top conventional ML algorithms, best-performing ML algorithms based on widely used performance measures, frequency and impact of language and type of data, pros and cons of standard and customized datasets, frequently used emotion sub-classes, and their impact.

To harness the valuable insights of industry experts, a survey employing the Analytical Hierarchy Process (AHP) was conducted, engaging 82 participants. AHP, a well-established methodology for multi-criteria decision-making, was chosen for its reputation for precision and accuracy in ranking and prioritizing the available options. The three criteria – Accuracy, Robustness, and Interpretability were considered for the evaluation of the five most frequently used machine learning algorithms. The survey not only draws upon the collective expertise of industry professionals but also facilitates a comprehensive assessment of the selected algorithms, contributing to the strength of the decision-making process in the domain of text-based emotion detection. A series of exhaustive experiments was also conducted to assess the effectiveness of the chosen ML algorithms. These experiments systematically investigated how text preprocessing, hyperparameter tuning, and different vectorization techniques influenced the performance of the selected models. Moreover, the research explored the complex concept of multi-level stacking by examining different ways the selected models could be combined.

2. Background

Emotion detection from text is a research area that has garnered a lot of interest from researchers around the world. Many studies have focused on different aspects of this topic, including the background theory of the various emotion models that have been used in literature and the different computational approaches that have been developed. Most surveys published on text emotion detection and analysis summarize the research conducted in this field, including existing emotion detection methods, approaches, datasets, experiments, and outcomes.

Studies like Acheampong et al. (2020), Murthy and Kumar (2021), and Nandwani and Verma (2021) have generally classified emotion models into two broad categories: Categorical or discrete models and dimensional models. Categorical models categorize emotions into a fixed number of categories, such as happy, sad, angry, and so on. Dimensional models, on the other hand, represent emotions as points in a multi-dimensional space, with each dimension representing a different aspect of the emotion. While some surveys, such as Alswaidan and Menai (2020) and Murthy and Kumar (2021) discussed various emotion models and classified them as appraisal models. An appraisal model of emotion proposes that emotions are the result of an individual’s evaluation or appraisal of a specific event or situation. This evaluation includes the person’s thoughts, beliefs, and expectations about the situation, as well as the contextual factors that may affect their emotional response. One of the most used emotion models is the Ekman model (Nandwani and Verma, 2021), which categorizes emotions into six basic categories: anger, surprise, disgust, joy, fear, and sadness.

However, none of the surveys have investigated the performance of different emotion models used in literature in comparison with each other. Moreover, in addition to the diverse emotion models employed, various computational methods have been developed for detecting emotions from textual data. These methods can be categorized into three main types: keyword-based, rule-based, and machine learning-based approaches. Keyword-based approaches utilize specific keywords or phrases to recognize emotions in text. Rule-based approaches employ a predefined set of rules to identify emotions in text.

Lastly, machine learning based approaches apply algorithms to learn patterns in training data and predict emotions in unseen data accordingly. While many studies have compared the different computational approaches for emotion detection from text, there has been less focus on the in-depth analysis of the performance of each approach. For example, it is not clear which machine learning algorithms perform best for emotion detection, selection of emotion models, or datasets and how they can impact the performance of a model. This lack of in-depth analysis is a gap in the existing research and motivates the need for a separate study specifically focused on conventional machine learning for text-based emotion detection. One reason why conventional machine learning algorithms may be particularly useful for emotion detection from text is that they do not require huge amounts of balanced data while on the other hand deep learning models require huge amount of balanced data for its training (Alswaidan and Menai, 2020) and most of the available dataset does not fulfill this requirement (Yuan and Purver, 2015). A balanced dataset is one where the different classes (in this case, emotion) are represented in approximately equal proportions. This is important because the model needs to be exposed to a representative sample of all the classes it will encounter to learn to make accurate predictions. In this systematic review, 55 articles were selected that used conventional machine learning algorithms and classified the algorithms according to their performance on different metrics such as F-Score, Precision, Recall, and Accuracy. We also investigated the impact of factors such as the type of dataset and the number of emotion classes on the performance of the algorithms.

The study also engaged 82 industry experts using the AHP, a survey known for its precision in multi-criteria decision-making. Focused on Accuracy, Robustness, and Interpretability, the research systematically compared the top five frequently used machine learning algorithms for emotion detection in text.

The study broadened its research methods by conducting thorough experiments to measure the effectiveness of specific machine learning algorithms. These experiments carefully investigated how the text pre-processing, tuning hyperparameters, and using different vectorization techniques affected how well the chosen models performed. Additionally, the research explored multi-level stacking, examining different ways the selected models could be combined for analysis.

3. Research methodology

This study uses an SLR to examine and summarize the existing literature. This SLR aims to determine the most effective machine learning algorithms, key factors, and their impact on the accuracy of emotion detection in text.

3.1. The SLR Process

According to Carrera-Rivera et al. (2022) and Paul and Barari (2022), SLR must contain three basic steps: planning, execution, and reporting. The initial steps are defining the study's goals and objectives, developing research questions, developing a search strategy for narrowing down the criteria for searching material relevant to the research questions, selecting the identified literature, and developing a data extraction strategy. The objective of this study is mentioned above, while the rest of the steps are described below.

3.2. Research questions

The Research Questions (RQ) are as follows:

1. Which algorithms are frequently used for text-based emotion detection?
2. Which algorithms outperform others?
3. What are the strengths of the best-performing algorithms, or reasons for their outperforming?
4. What types of datasets are mostly used for text-based emotion detection? i.e., Standard or Customized, and why?
5. What is the Impact of Emotion Classes on the Performance of the selected ML algorithm?

Text-based emotion detection is a challenging task due to the complexity of human emotions and the subtleties of language use. Therefore, different approaches have been developed and used in the literature to address this challenge. The purpose of this study is to analyze conventional machine learning models only. RQ1 and RQ2 can help researchers and practitioners to know the most appropriate algorithm. Different algorithms may perform better in different contexts, so understanding which algorithms are frequently used and which ones are most effective can help inform decision-making. Comparing the performance of different algorithms can provide insights into their strengths and weaknesses and identify which algorithms are most effective for a given task. Similarly, researchers can focus on improving the performance of the most effective algorithms or developing new algorithms that can overcome the limitations of current ones.

The motivation of RQ3 is to provide researchers with insights into the features, models, and techniques that are most effective for capturing the emotional content of text data. By identifying the strengths and weaknesses of different algorithms, researchers can tailor their models to take advantage of the strengths of the most effective techniques and avoid the limitations of less successful models. Researchers have used a variety of datasets, including standard datasets as well as customized datasets. Standard datasets are used due to their free and public availability, facilitating comparison across studies and allowing for replication and validation of results.

Customized datasets, on the other hand, are often created for a specific research question or domain and may include emotions that are not included in standard datasets. Understanding the types of datasets used is important because the quality and representativeness of the data can significantly impact the accuracy and reliability of the emotion detection method, which is the motivation of RQ4. RQ5 helps in understanding the impact of emotion classes on the performance of the selected machine learning algorithm. By understanding the impact of emotion classes on performance, researchers can make informed decisions about the optimal choice of emotion classes, given the available resources and the specific needs of the problem.

3.3. Search strategy

This study conducted a comprehensive search of published literature in academic journals and conferences, utilizing the following digital libraries:

- Google Scholar
- Science Direct
- IEEE Explore
- ACM
- Emerald
- Springer Link
- Scopus

3.4. Search terms

To construct the most effective search string, we derive major terms from the objective and research questions, and identify alternative terms and synonyms for each term. Then use Boolean OR to incorporate alternatives and synonyms, and Boolean AND to link major terms i.e., (“Emotion Detection” OR “Emotion Identification” OR “Emotion Recognition”) AND (“Text Mining” OR “Text Data Mining” OR “Text”) AND (“Machine Learning”).

The seven electronic databases were searched for journal articles and conference papers using the search terms that were created above. Since the search engines of various databases use various search string syntaxes, the search terms were modified to accommodate multiple databases.

3.5. Search process and selection

SLR necessitates a thorough investigation into relevant literature. To discern the most relevant publications, we adopt the tollgate approach, proposed by Kumar et al. (2023). Initially, publications undergo scrutiny based on their titles, abstracts, and keywords, by comparison with our research questions. Afterward, the selected studies underwent additional review based on their ’Introduction’ and ’Conclusion’ sections. Then, the complete texts of these sources are scrutinized for further refinement, employing inclusion and exclusion criteria. Moreover, the chosen studies undergo cross-examination by a secondary reviewer to uphold the transparency of the process. The entire procedure is documented to provide justification for the inclusion or exclusion of studies in the final evaluation, as shown in Table 1.

1. The process involved conducting separate searches across seven electronic databases, with the resultant papers collated to establish a set of candidate papers. During this phase, a total of 4,079 studies were identified as candidate papers.
2. The candidate studies were reviewed based on ‘Title and Abstract’ in this phase, which returned 1,011 relevant studies.
3. In this phase, the selected studies were further reviewed based on ‘Introduction and Conclusion,’ which reduced the candidate papers to 264.
4. 97 studies were selected in this phase after a full-text review.
5. Finally, after removing duplicates and applying Quality Assessment Criteria, 55 studies were selected for data extraction.

3.6. Inclusion criteria

- The study pertaining to the search terms outlined in the preceding section was incorporated.
- For studies that have both conference and journal versions, only the journal version was considered for inclusion.
- In cases of duplicate publications pertaining to the same study, only the most comprehensive and recent version was selected for inclusion.
- Studies that were in English were included.

3.7. Exclusion criteria

- Studies that did not focus explicitly on emotion detection in text through ML were excluded.
- Duplicate studies were excluded.
- Studies that were in a language other than English were excluded.

3.8. Quality assessment criteria

Following the application of the quality assessment criterion to the results, 55 studies were chosen for inclusion in the final list. The primary objective of the quality assessment is to scrutinize and evaluate the nature of the ultimately selected papers. The quality checklist comprises the following questions:

- Is there sufficient data available to substantiate the results?
- Does the researcher convey a tendency to emphasize reporting positive results over negative ones?

- Is the objective of the research clearly articulated and unambiguously defined?
- Are the research outcomes connected to the stated objective of the study?
- Is the discussion of the Emotion Detection context clearly and thoroughly presented?

3.9. Data extraction

Following the extraction of data, no disagreements were identified. The subsequent data was extracted from each of the selected scientific articles, summary of the selected articles for this SLR is also presented in Table 2.

- Title of the paper
- Authors
- References
- ML Algorithms used
- Outperforming ML algorithms
- Performance measuring parameters, i.e., Accuracy, Precision, F-measure, and Recall.
- Number and Size of Datasets used
- Type of Datasets used, i.e., Standard or Customized
- Types of data utilized, i.e., Customer reviews, social media posts, emails, news headlines, etc.
- Dataset Language
- Number of Emotion sub-classes

Table 1: Summary of the five phases of the tollgate approach

Phase	Details
Phase 1	Total Candidate Studies = 4079 (Google Scholar: 2410, Science Direct: 118, IEEE Explore: 215, ACM: 53, Emerald: 571, Springer: 186, Scopus: 526)
Phase 2	Studies selection based on Title and Abstract = 1011
Phase 3	Study selection based on Introduction and Conclusion = 264
Phase 4	Studies selection based on Full text review = 97
Phase 5	Studies selected for data extraction = 55

4. Findings of the review

4.1. RQ1: Which ML algorithms are frequently used for text-based emotion detection?

As shown in Fig. 1, SVM is the most frequently used ML algorithm. 72% of the selected studies have used SVM, and 56% of the selected studies have used Naïve Bayes (NB), which is the second most frequently used algorithm for text-based emotion detection. While Random Forest (RF), Decision Tree (DT), K-Nearest Neighbor (KNN), and Logistic Regression (LR) have percentage usage of 32%, 27%, 23% and 20% respectively. Ensemble classifiers like bagging, boosting, and stacking were also used. 47% of the selected studies also used algorithms that are represented by ‘Other’ in Fig. 1.

4.2. RQ2: Which machine learning algorithms outperform others?

The accuracy of recognizing an emotion is reliant on the selection of a suitable machine learning algorithm and its hyperparameter optimization, which is nothing but choosing a set of optimal hyperparameters for a learning algorithm. Although the process of training a machine learning algorithm is the same, every algorithm has its unique characteristics. For this reason, most of the researchers train a set of relevant algorithms and, by comparing their performance, they find the most suitable algorithm. Fig. 2 shows the most frequently outperforming machine learning algorithms. SVM is the most frequent and popular best-performing machine learning algorithm.

In our case, 21 studies (38% of the selected studies) consider SVM as the best-performing algorithm for text-based emotion detection. Naïve Bayes is the 2nd best performing algorithm reported by 14% of the selected studies. Fig. 2 shows the top six best-performing algorithms with their frequencies in the selected studies. Fig. 3 shows the Accuracy, F-score, Precision, and Recall of the top-performing ML algorithms. It can be seen clearly from Fig. 3 that SVM, NB are consistently performing well with respect to Accuracy, F-Score, Precision, and Recall, while LR and XGB are competing with them and are even better in some respects, but very low in frequency as compared to SVM and NB. Other machine learning like Random Forest (RF), Decision Tree (DT), have inconsistent performance.

4.3. RQ3: What are the strengths of best best-performing algorithms or reasons for their outperforming?

Every classifier has its own unique strengths as well as limitations, which help us select the appropriate one among others. In this case, 38% of the selected studies declare SVM the best-performing algorithm. This is since SVM can handle non-linear decision boundaries and capture complex relationships between the features and the target variable using a kernel function, which is especially useful when working with high-dimensional text data. A kernel function is a mathematical function that maps the input data into a higher-dimensional feature space, where it may be more separable. The SVM then finds the maximum-margin hyperplane in this transformed feature space, which corresponds to a non-linear decision boundary in the original feature space. The kernel function allows SVMs to implicitly capture complex relationships between the features and the target variable, without the need to specify the mapping to the higher-dimensional feature space explicitly. This is because the kernel function computes the dot product of the feature vectors in the transformed feature space, without computing the transformation itself (Kalcheva et al., 2020).

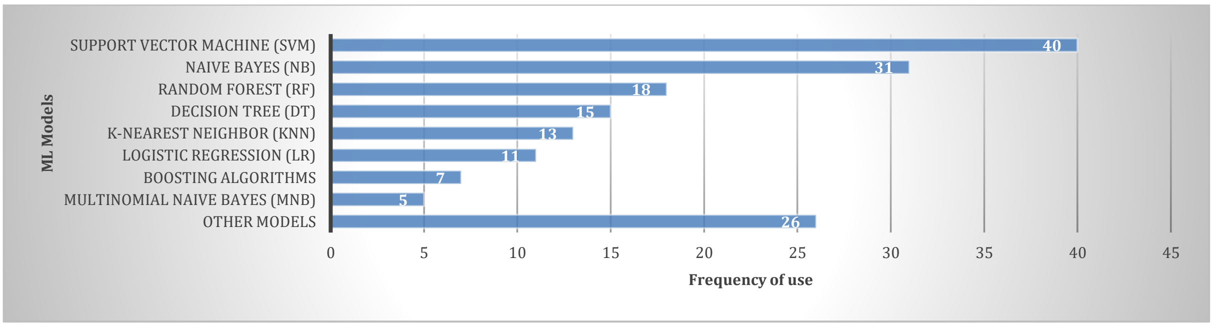


Fig. 1: List of frequently used ML algorithms for emotion detection in text

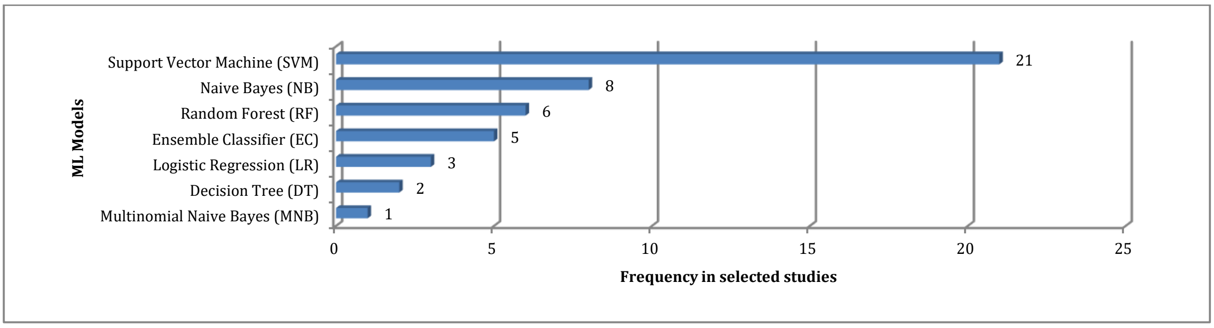


Fig. 2: Frequency of top-performing ML algorithms in emotion detection

DT mostly suffers from overfitting, particularly when dealing with high-dimensional data like text (Halim et al., 2020). On the other hand, RF is a type of ensemble learning method that builds multiple decision trees and averages their outputs, which may not be as effective in capturing complex relationships between features as SVM (Chakriswaran et al., 2019). SVM can handle class imbalance better because SVM is inherently a binary classifier, which means they are optimized for finding the hyperplane that separates the two classes with the maximum margin. This is specifically useful in situations where the minority class is important and requires more attention, as SVM will focus on correctly classifying instances of these minority classes. Similarly, SVM has a regularization parameter that can help prevent overfitting and is less susceptible to the "curse of dimensionality," while DT and RF may suffer from overfitting, which may affect their performance.

In text data, there are often many words that may not contribute significantly to emotion prediction. Naive Bayes tends to be less affected by irrelevant features, making it more robust in the presence of noisy data. Naive Bayes and XGBoost can handle sparse data well, and they perform reasonably well with a high-dimensional feature space.

4.4. RQ4: What types of datasets are mostly used for text-based emotion detection? Standard or customized and why?

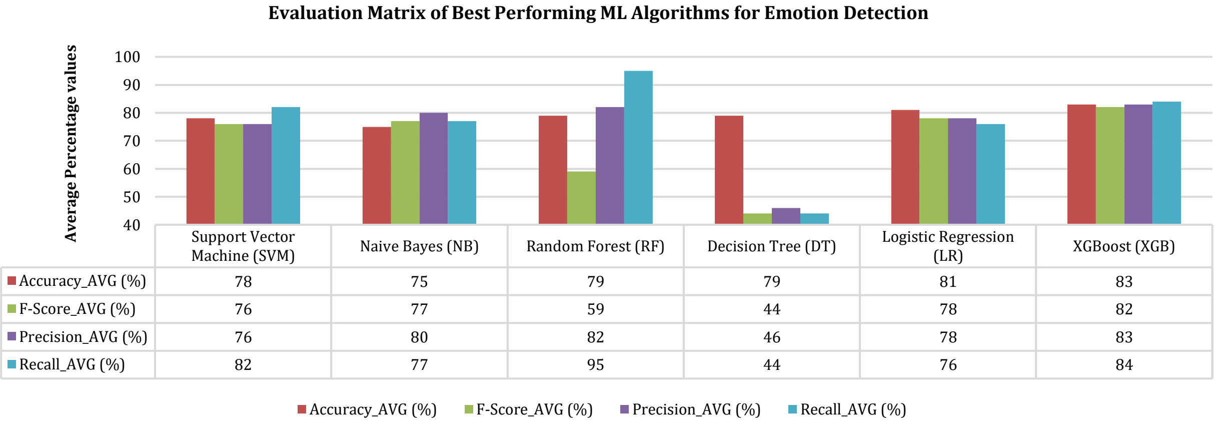
As shown in Fig. 4, 80% of the selected studies prefer to use their own customized dataset for text-based emotion detection despite the fact that data collection and annotation are time-consuming processes and require a lot of manual efforts. Although there are tools available that automate the process of annotation, their quality is still not reliable. The ability to accurately identify an emotion depends on the size, quality, and balance of the dataset. According to Alswaidan and Menai (2020), most of the available datasets, except ISEAR, are imbalanced.

4.5. RQ5: What is the impact of emotion classes on the performance of the selected ML algorithm?

For this study, we have considered only those publications that have datasets consisting of at least four emotion classes. Although some studies consider the high number of emotion classes like (Sintsova et al., 2014) consider about 20 emotional classes, few studies consider more than one dataset having different number of emotion classes each, like (Almahdawi and Teahan, 2018), but the most frequently used classes range from 4 to 8. Fig. 5 shows the frequencies of emotional classes. Most studies consider the Ekman model for emotion detection (Nandwani and Verma, 2021), consisting of 6 emotion classes used by 29% of the selected studies, while the 4 emotion classes were used by 21% of the selected studies. From the literature, it is recommended to use 4 or 8 emotion classes to achieve consistent results.

5. Analytical hierarchy process

The AHP, introduced by Abdolvand et al. (2015), is a renowned technique employed for making decisions involving multiple criteria. AHP is a method that assists in addressing intricate decision-making problems encompassing both quantitative and qualitative factors. Researchers from diverse disciplines have extensively investigated and applied AHP (Abrar et al., 2023a; 2023b; Kabra et al., 2015; Krenicky et al., 2022). Additionally, the AHP is utilized to prioritize machine learning algorithms in text-based emotion detection, considering their relative significance. The AHP comprises two phases: structuring the hierarchy and priority setting through pairwise comparisons.



	Support Vector Machine (SVM)	Naive Bayes (NB)	Random Forest (RF)	Decision Tree (DT)	Logistic Regression (LR)	XGBoost (XGB)
■ Accuracy_AVG (%)	78	75	79	79	81	83
■ F-Score_AVG (%)	76	77	59	44	78	82
■ Precision_AVG (%)	76	80	82	46	78	83
■ Recall_AVG (%)	82	77	95	44	76	84

Fig. 3: Best performing ML algorithms with average performance parameter

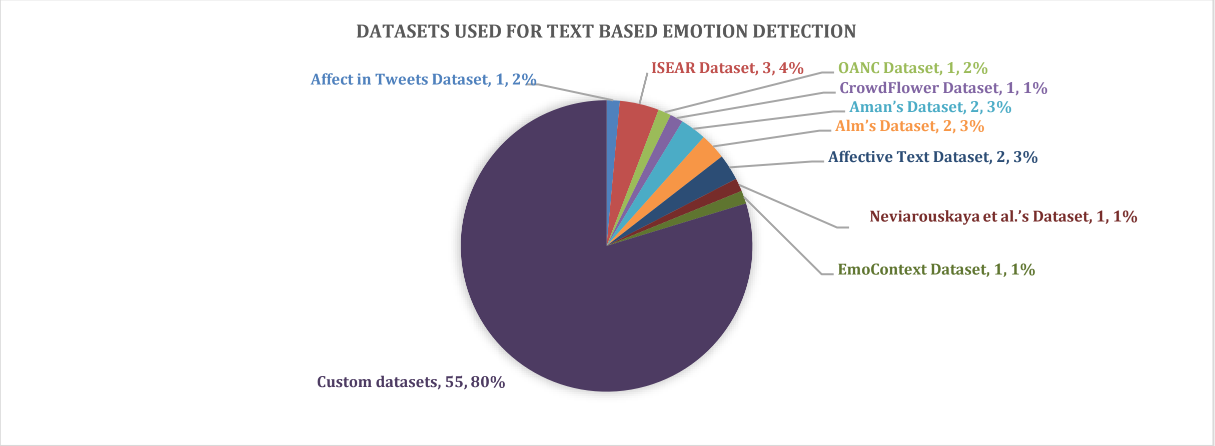


Fig. 4: Summary of datasets used for text-based emotion detection

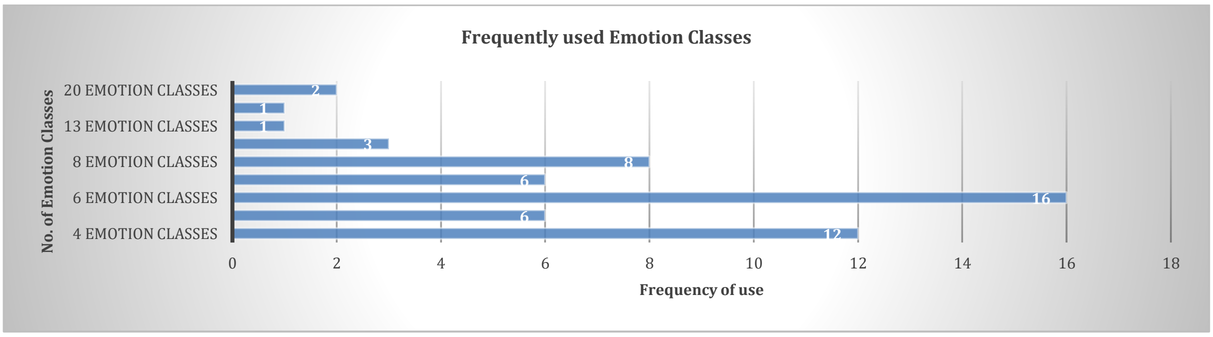


Fig. 5: Average accuracy of frequently used emotion classes

5.1. Structuring the hierarchy

In our research, we utilized the SLR to ascertain the machine learning algorithms predominantly employed and demonstrate superior performance in text-based emotion detection. Through the SLR process, we identified a set of five algorithms that were most utilized and yielded the best results. During Phase 1, experts were involved in the selection of three criteria, namely accuracy, robustness, and interpretability, from a range of available criteria such as accuracy, robustness, interpretability, time complexity, scalability, resource requirements, and model complexity. Accuracy refers to the algorithm's capability to accurately predict or classify unseen data, with higher accuracy indicating superior performance. Robustness evaluates how well the algorithm performs when confronted with noisy or incomplete data, outliers, or adversarial attacks. A robust algorithm should maintain its performance even under such challenging circumstances. On the other hand, interpretability pertains to the algorithm's capacity to provide an understanding and explanation of its decision-making process, aiding in the identification of internal biases specific to certain types of data.

Table 2: Summary of selected studies on ML algorithms for emotion detection with key features, strengths, limitations, and performance metrics

Reference	Algorithm	Outperforming algorithm	Features	Data	Ensemble classifier	Strengths	Limitations	Performance
Plaza-del-Arco et al. (2020)	SVM LR NB	SVM	N-Grams TF Emotional intensity scores	Tweets	4	Incorporation of affective features improved classification results. Focus on the Spanish language enhances contributions to this less-studied language in NLP.	Challenges with translations affecting contextual accuracy. Limited improvement observed with external knowledge bases.	F1 score: 71%
Zhang et al. (2016)	KTM SVM	KTM	N-Grams POS tagging Pointwise Mutual Information (PMI)	Blog Posts	19	Employs a hierarchical classification structure, allowing the model to handle different levels of granularity in emotion classification.	Limited generalizability. The proposed model is computationally intensive and complex to implement.	F1 score: 83%
Tuhin et al. (2019)	NB TA	TA	TF-IDF Emotion class probabilities	Sentences	6	Addresses a significant gap in sentiment analysis research for the Bangla language. Demonstrates high accuracy levels, which shows effectiveness over traditional methods like NB in both sentence and article-level analyses. Explores sentiment analysis at multiple levels, increasing the applicability of the research in real-world scenarios.	Limited ability of the model to generalize beyond the scope of the training data provided. The study focuses on a limited number of emotional classes, which might not capture the full spectrum of human emotions in text. Does not discuss the computational efficiency or scalability of the proposed methods.	Accuracy: 90%
Kang et al. (2017)	DWET HDWET	HDWET	-	Blog Posts	8	Enhance the granularity of emotional understanding. Comparison of the proposed model against traditional emotion recognition algorithms.	The performance of the proposed model heavily relies on the quality and extent of emotion annotation in the training dataset. Language specificity: Focused on Chinese text, which may limit the direct applicability of findings to other languages without modifications.	F1 score: 53%
Suhasini and Srinivasu (2020)	ME SVM	ME	N-Grams POS tagging	Tweets	5	Applied a two-stage classification process that effectively separates emotional from non-emotional tweets before classifying them into specific emotion types. Comprehensive Feature Utilization: Included a variety of features to enhance model accuracy.	The detailed tuning and model complexity might lead to overfitting. Limited Emotion classes were considered. High dependence on crafted features, which may limit the ability to adapt the model to other domains or languages without significant re-engineering.	Accuracy: 72%
Povoda et al. (2015)	SVM	SVM	TF-IDF	Helpdesk Messages	5	Innovative Approach to automate emotion recognition in text-based communications. Proposal for language-independent methods that could be adapted for non-English datasets.	Relatively small dataset. Language-specific model.	Accuracy: 77%
Gunarathne et al. (2013)	SVM	SVM	TF-IDF-CF	Instant Messages	6	Improved communication effectiveness through real-time processing. User-friendly, intuitive, and engaging interface.	Focused on a limited, predefined set of emotions. Dependence on Predefined Corpus. The application processes sensitive personal data can raise privacy issues if not handled properly.	Accuracy: 78%

Ghanbari-Adivi and Mosleh (2019)	DT
	KNN
	RF
	AB
	GB
	CNN
	MLP
	EC
		EC	Doc2Vector
			Dependency Parsing
				2 Datasets: Sentences;
				1 Dataset: Tweets;
					6	Use a sophisticated parameter tuning approach. Employing both traditional and irregular text datasets to enhance model applicability.	The proposed model has a high computational cost and complexity. May not generalize well outside the specific datasets or emotions studied without further validation. Dependency on accurate pre-processing and feature extraction might limit its application in less structured text.	Regular Sentences Accuracy: 99.49%
								Irregular Sentences Accuracy: 88.49%

Pang et al. (2019)	SVM	SVM	BoW POS tagging	Blog Posts	7	Detailed analysis of text data in a language with complex semantic structures. Extensive experimentation with both document-level and sentence-level emotion classification.	Dependence on labeled data from a single source. Possible overfitting to specific idiomatic expressions. Imbalance in the dataset distribution of emotions.	Recall: 73%

Almahdawi and Teahan (2018)	PPM
	NB
	SMO
		PPM	Vectorization	2 Datasets: Blog Posts;
				1 Dataset: Fairy Tales;
					6	Demonstrated effectiveness across various types of text data. High accuracy in classifying both broad categories of emotion and specific emotions.	Limited evaluation of linguistic diversity. Potential overfitting to specific dataset characteristics due to the high specialization. The impact of cross-linguistic and cultural factors in emotion recognition is missing.	Accuracy: 88%

Tian et al. (2014)	SVM NB LB RF	RF	POS tagging	Sentences	5	Development of a case-based reasoning system for emotion regulation. Utilization of a diverse array of ML algorithms to compare effectiveness. Focus on a unique dataset of interactive Chinese texts, which is less common in emotion recognition research.	Limited to specific types of interactive texts. Cultural biases in emotion perception.	Recall: 56%
Jain et al. (2017)	NB SVM	SVM	-	Tweets	6	Detailed empirical analysis across multiple real-world domains. Use publicly available resources to enhance feature sets for emotion classification.	Relies on existing lexical resources, which might not cover all nuances of emotional expression in multilingual settings.	Precision: 86%
Halim et al. (2020)	RF SVM LR KNN	RF	BoW N-Grams TF-IDF	Emails	4	Application to professional communication scenarios enhances real-world relevance. Thorough comparison of multiple ML algorithms on a consistent dataset.	Limited generalizability due to dataset specificity. Dependence on crafted features, which might not transfer well across different text types or languages.	F1 score: 75%
Ray et al. (2021)	NB	NB	Sentiment Scores	4 Datasets: Customer Reviews	8	Utilization of real-world data from multiple domains. Significant improvement of the combined approach compared to traditional models.	Performance is highly dependent on the quality of the collected social media data. Analysis of decimal ratings or emoticons is missing, which could be relevant in social media contexts.	Accuracy: 58.34% Accuracy: 100% Accuracy:

								57.84%
Povoda et al. (2016)	SVM KNN RF	SVM	N-Grams Lemmatization Synonym Replacement	Helpdesk Messages	5	Utilization of a large and diverse dataset. Methodology is adaptable to other languages, enhancing the system's versatility.	Dependence on manually labeled data may introduce human bias and require extensive resources to prepare.	Accuracy: 86.89%

Patil and Patil (2013)	SVM
	DT
	EC
		EC	BoW
			Affective Words
				1 Dataset: Blog Posts;
				1 Dataset: Tweets;
					4	Improved emotion classification accuracy. Enriched user interaction by generating visual emotional content.	Reliance on manually annotated data limits scalability and real-time applications.	Accuracy: 89.38%

Balakrishnan and Kaur (2019)	MNB NB	MNB	BoW	Facebook Posts	8	Utilization of a real-world dataset reflecting genuine user interactions. Handling of real social media text complexities, such as slang and non-standard expressions.	The performance metrics could vary significantly for outsiders of the diabetes community. Limited comparison with other machine learning models.	F-Score: 82%
Kaur et al. (2020)	NB SVM	SVM	Normalization	Survey	8	Utilization of machine learning to quantify and classify complex emotional data from textual feedback. Potential to enhance teaching and learning strategies based on detailed student feedback analysis.	Limited to feedback from a single course, which may not be generalizable across different courses or academic disciplines. Possible bias in student responses based on their individual experiences and perceptions, which may not reflect the overall effectiveness of the teaching methods.	F-Score: 85%
Saad et al. (2018)	SVM DT	DT	TF-IDF	Folklores	4	Provides a foundational study for integrating emotional recognition into storytelling speech synthesis.	Substantial room for improvement in the overall results of emotion recognition. Limited comparison with other machine learning models.	Accuracy: 62.5%
Gohil and Patel (2019)	LR MNB SVM	SVM	TF-IDF	Tweets	8	Robust methodology for emotion analysis across multiple languages. Contribution to the rare field of multilingual emotion detection.	Dependency on accurate translation for feature generation.	F-Score: 84%
Patacsil (2020)	NB KNN EC	EC	N-Grams	Blog Comments	7	Innovative use of automatically generated datasets for emotion classification. Incorporation of language translation to expand dataset usability across languages.	The dataset was skewed towards certain emotions, which could bias the model's performance.	Accuracy: 76%
Esmin et al. (2012)	SVM NB	SVM	N-Grams TF-IDF	Tweets	6	Utilizes a novel hierarchical classification approach that improves classification performance. Employs a real-world dataset that enhances the practical applicability of the findings.	Reliance on manual annotation for training data can introduce subjective biases and may not scale well.	Accuracy: 80.35% Precision: 69% Recall: 91% F-Score: 80%
Sintsova et al. (2014)	BWV NB PMI	BWV	N-Grams	Tweets	20	The weighted voting approach significantly improved classification accuracy.	Imbalance in emotion distribution in the dataset can affect the learning process and the model's ability to generalize.	
Putra et al. (2020)	KNN RF NB LR SVM	SVM	TF-IDF BoW N-Grams	Tweets	4	Use of expert validation to ensure the accuracy of emotion annotation in the dataset.	Limited emotion classes are considered. Potential biases due to the selection of specific hashtags for data collection.	Accuracy: 95%
Liu and Qi (2018)	LR NB KNN RF SVM	LR	VSM Chi-Square Values	Customer Reviews	6	Tailored to the nuances of Chinese linguistic and emotional expressions. Utilizes a comprehensive and culturally relevant dataset from a major e-commerce platform.	Reliance on manual annotation for training data.	F1 score: 76%
Chowanda et al. (2021)	NB GLM ANN DT RF SVM	GLM	N-Grams TF TO TF-IDF	Tweets	4	Effective use of a large and diverse dataset for training and testing the models.	Potential bias due to imbalance in the original dataset before sampling adjustments. Potentially overfit for DT and Random Forest.	Accuracy: 90.2%
Hussein et al. (2020)	NB KNN SVM	NB	N-Grams POS tagging TF-IDF	1 Dataset: Social Media Posts; 1 Dataset: Blog Posts	4	Utilization of real-world data sources. Detailed methodology for preprocessing and feature selection.	The paper lacks discussion on the impact of imbalanced classes or the handling of sarcasm and idioms. No discussion on computational efficiency or scalability of the proposed methods.	Accuracy: 70%
Sarakit et al. (2015)	MNB DT SVM	SVM	TF TF-IDF	YouTube comments	6	Provide a broad view of potential methodologies for emotion classification. Utilizes real-world data.	The ambiguity in comments and use of slang might lead to misclassification and negatively affect accuracy.	Accuracy: 76.14%
Mahajan and Zaveri (2021)	SVM MLP REPT DT	REPT	-	Text Conversations	4	Utilizes a diverse set of features for robust emotion recognition. Achieves competitive performance with traditional less resource-intensive machine learning models.	The model may require adjustments to deal with nuanced emotional expressions due to the simplicity of the features used. Relies on a specific set of labeled data, which may not generalize across other datasets.	Accuracy: 75.88%
Saputri et al. (2018)	LR SVM RF	LR	BoW Word2Vec POS tagging	Tweets	5	Comprehensive feature engineering to identify the best features for emotion classification in Indonesian tweets.	Not consider multi-label emotion classification. Not generalizable for long text features.	F1 score: 69.73%
Nguyen and	MLR	MLR	CV	Facebook	6	Demonstrates structured methodology and impactful techniques for data	Certain preprocessing techniques might remove some contextual	F1 score:

Van Nguyen (2020)	CNN		TF-IDF	comments		preprocessing.	nuances, critical to understanding certain emotions. Reliance on manually annotated datasets may introduce bias or errors based on the annotators' interpretations.	64.40%
Balakrishnan et al. (2021)	SVM RF NB	RF	TF-IDF	Customer Reviews	6	Employs various statistical methods to ensure robustness in results.	Reviews from a limited audience from a specific period were only analyzed.	F1 score: 58.8%
Alotaibi (2019)	LR	LR	TF-IDF	Sentences	7	A well-defined framework enhances the reproducibility of the research.	No comparison with other machine learning models.	Precision 86% Recall 84% F1 score: 85%
Majeed et al. (2020)	KNN DT SVM RF	SVM	Word2Vec	Social Media Posts	6	Pioneering work on emotion detection in Roman Urdu, addressing a significant gap in language processing for underrepresented languages. Development of a large, diverse corpus specifically for Roman Urdu, enhancing research capabilities in this area.	Comparison of different vectorization techniques is missing.	F1 score: 69% Precision: 70% Recall: 70% Accuracy: 69.54%
Mondal and Gokhale (2020)	RF SVM MLP NB GB	RF SVM	N-Grams TF-IDF	Tweets	8	Robust Dataset Utilization. Provides extensive metrics on model performance.	Dependence on Manual Annotation.	Accuracy: 78.5%

Parvin and Hoque (2021)	LR
	MNB
	SVM
	RF
	DT
	KNN
	AB
	EC
		EC	BoW
			TF-IDF
				2 Datasets: Social Media Posts;
				1 Dataset: Customer Reviews;
					6	Utilization of ensemble methods to improve classification accuracy.	Comparison of different vectorization techniques is missing.	F1 score: 62.39%

Chaffar and Inkpen (2011)	SVM
	NB
	DT
		SVM	BOW
			N-Grams
				1 Dataset: News Headlines;
				1 Dataset: Fairy Tales;
				1 Dataset: Blog Posts;
					6	Diverse and heterogeneous datasets are utilized, enhancing the generalizability of the findings.	Complicate the modeling process.	Accuracy: 81.16%

Sreeja and Mahalakshmi (2019)	SVM LR NB PEREM	PEREM	POS tagging TF-IDF	Poetry	9	Creation of a specialized corpus for poetry, which is made publicly available. Use of a comprehensive set of emotions to reflect a wide range of human sentiments in literature.	Limited to the emotions defined in 'Navarasa,' which might not encompass all possible emotional expressions in poetry.	Precision: 88% Recall: 86% F-measure: 87%
Angel Deborah et al. (2020)	RF AB GB	GB	POS tagging BoW	Text	4	Focus on contextual understanding of emotions in text, which is a challenging aspect of natural language processing. Practical implications for enhancing communication interfaces with emotional intelligence.	The bias towards the "others" category in the dataset could affect the generalizability of the models. Reliance on a single dataset might limit the application scope to similar text contexts.	Accuracy: 85.98% F-score: 86% Precision: 86% Recall: 86%
Abdullah et al. (2020)	NB SVM DT	NB	N-Grams TF-IDF	Tweets	4	Effective handling and preprocessing of Arabic text, which It is complex due to its script and grammar.	Limited to only four emotion classes.	Accuracy: 80.1%
Sailunaz and Alhajj (2019)	NB SVM KNN DT	NB	-	Tweets	7	Utilizes real-time data from social media. High-performance metrics indicate strong model accuracy.	Small dataset of 500 tweets. The study does not explore the impact of context or sarcasm in tweets.	Precision: 98.1% Recall: 98.5% Accuracy: 93.1%
Suhasini and Srinivasu (2020)	NB KNN	NB	-	Tweets	4	Leveraged an existing, large-scale public dataset, facilitating reproducibility and scalability.	Emotion classification is somewhat basic, focusing on binary dimensions of emotion rather than a more nuanced spectrum.	Accuracy: 72.6% Precision: 73% Recall: 73% F-Score: 72.5%

5.2. Priority setting through pairwise comparisons

The AHP, a decision-making technique introduced by Kabra et al. (2015), is utilized in scenarios involving multiple criteria. AHP has been validated as an effective solution for complex decision-making challenges across diverse research areas (Abrar et al., 2023b). Our goal is to conduct a priority-based ranking of machine learning algorithms in the context of text-based emotion detection. The implementation steps of AHP are depicted in Fig. 6. The following sections provide an explanation of the steps illustrated in Fig. 6.

5.3. Decompose a complex issue into a structured hierarchy

This step involves identifying objectives and criteria, along with arranging machine learning algorithms by their importance. The issue is organized in a hierarchical manner across at least three levels, as shown in Fig. 7. The top level (level 1) outlines the primary objective of the problem. The criteria and options are laid out at levels 2 and 3, respectively.

5.4. Construction of pair-wise matrices

To prioritize the ML algorithms and their criterion using AHP, a pairwise comparison survey was conducted. The survey was conducted by 82 participants, which might lead to questions about how well the sample represents the wider population and how it could affect the study's results. However, it is important to note that the AHP is designed to work well even with smaller groups of participants (Krenicky et al., 2022). This method has been used successfully in past research, even with limited sample sizes. For instance, Shameem et al. (2018) and Saravia et al. (2018) used small groups of just five and nine participants, respectively, to study opinions, experiences, or to determine the importance of different factors. Similarly, Pang et al. (2019) surveyed intelligent building systems using the AHP method with a sample of nine experts. Based on these examples, the sample size of 82 responses in this study seems sufficient for analyzing the data gathered using the AHP method. To ascertain the relative significance of the ML algorithms and their criteria, we utilized a consistent approach of constructing pairwise comparison matrices for each criterion. A standardized 7-point comparison scale, presented in Table 3 and explained in Table 4, was employed to evaluate the importance of each ML algorithm and criterion.

Table 3: Details of the intensity scale

Description	Significance intensity
Equally important	1
Moderately important	3
Strongly more important	5
Very strongly more important	7
Intermediate values	2,4,6,8

5.5. Calculate the priority weight of the ML algorithms

To determine the priority weights of the Algorithm, a pair-wise comparison is conducted (Gohil and Patel, 2019). ML Algorithms are compared at each level based on their relative importance and the criteria defined at the higher levels. The pair-wise comparison matrices are utilized to calculate the priority weight using the following approach.

Matrix: A comparison matrix for ML algorithms, focusing on pairwise evaluations. The discussion of these pairwise matrices can be found in the "Construction of Pairwise Matrices" section.

Normalization: Normalization of the matrix involves dividing each value in every column by the sum of that respective column. The pairwise matrices presented in Section "Construction of Pairwise Matrix" as Tables 5-8 undergo further processing in which each value is divided by the sum of its column. This subsequent step results in the generation of normalized matrices shown in Tables 9-12.

Priority Weight: Determine the average of each row in a matrix as part of the normalization process. Eq. 1 calculates λ_{max}, Eq. 2 calculates CI while Eq. 3 calculates CR for the criteria given in Table 9.

$$\lambda_{max} = (1.8262 \times 0.5407) + (4.1571 \times 0.2671) + (5.0296 \times 0.1922) = 3.0644 \quad (1)$$

$$CI = \frac{Equation\,(1)-3}{3-1} = 0.0322 \quad (2)$$

$$CR = \frac{Equation\,(2)}{0.58} = 0.055491781 \leq 0.1 \ (Consistency\ Okay) \quad (3)$$

Eq. 4 calculates λ_{max}, Eq. 5 calculates CI while Eq. 6 calculates CR for the Normalized Matrix based on Accuracy, given in Table 10.

$$\lambda_{max} = (3.3225491 \times 0.3051381) + (5.8895882 \times 0.1864840) + (5.6101952 \times 0.1829261) + (4.8455310 \times 0.2035673) + (8.0033319 \times 0.1218843) = 5.10027489 \quad (4)$$

$$CI = \frac{Equation\,(4)-5}{5-1} = 0.025068725 \quad (5)$$

$$CR = \frac{Equation\,(5)}{1.12} = 0.02238279 \leq 0.1 \ (Consistency\ Okay) \quad (6)$$

Eq. 7 calculates λ_{max}, Eq. 8 calculates CI while Eq. 9 calculates CR for Normalized Matrix based on Robustness, given in Table 11.

$$\lambda_{max} = (3.2301 \times 0.3116) + (5.4474 \times 0.2103) + (5.7571 \times 0.1789) + (5.6905 \times 0.1765) + (7.9363 \times 0.1227) = 5.1601 \quad (7)$$

$$CI = \frac{Equation\,(7)-5}{5-1} = 0.04001601 \quad (8)$$

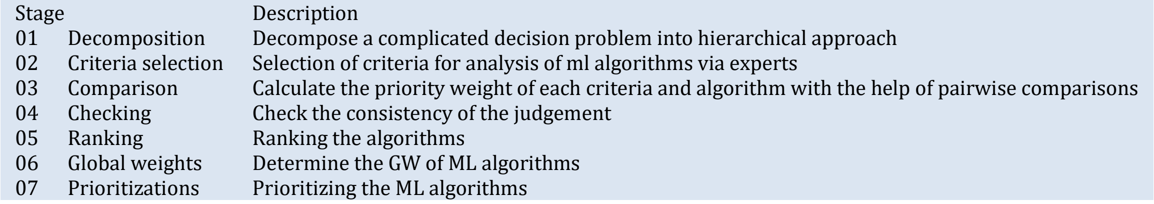
$$CR = \frac{Equation\,(8)}{1.12} = 0.0357 \leq 0.1 \ (Consistency\ Okay)\ 1601 \quad (9)$$

Eq. 10 calculates λ_{max}, Eq. 11 calculates CI while Eq. 12 calculates CR for Normalized Matrix based on Interpretability, given in Table 12.

$$\lambda_{max} = (3.6276 \times 0.2745) + (4.6600 \times 0.2348) + (5.4531 \times 0.1920) + (6.6579 \times 0.1531) + (6.7079 \times 0.1456) = 5.1329 \quad (10)$$

$$CI = \frac{Equation\,(10)-5}{5-1} = 0.0332 \quad (11)$$

$$CR = \frac{Equation\,(11)}{1.12} = 0.0297 \leq 0.1 \ (Consistency\ Okay) \quad (12)$$



Stage		Description
01	Decomposition	Decompose a complicated decision problem into hierarchical approach
02	Criteria selection	Selection of criteria for analysis of ml algorithms via experts
03	Comparison	Calculate the priority weight of each criteria and algorithm with the help of pairwise comparisons
04	Checking	Check the consistency of the judgement
05	Ranking	Ranking the algorithms
06	Global weights	Determine the GW of ML algorithms
07	Prioritizations	Prioritizing the ML algorithms

Fig. 6: AHP stages

Table 4: Description of the 9 Likert scale for the intensity of the importance

Size of matrix	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.9	1.12	1.32	1.41	1.45	1.45	1.49

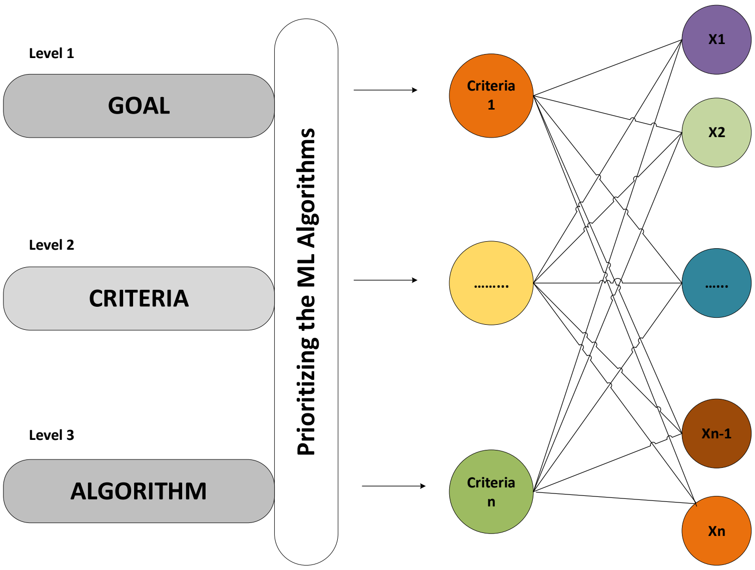


Fig. 7: The hierarchical structure of the prioritization process

Table 5: Matrix for pair-wise comparison of criteria

	Accuracy	Robustness	Interpretable
Accuracy	1	2.587	2.275
Robustness	0.387	1	1.755
Interpretable	0.440	0.570	1

Table 6: Pair-wise matrix of alternatives based on accuracy

	SVM	NB	DT	RF	LR
SVM	1	2.506	1.762	1.216	1.875
NB	0.399	1	1.351	0.989	1.582
DT	0.567	0.740	1	1.097	1.709
RF	0.823	1.011	0.911	1	1.838
LR	0.533	0.632	0.585	0.544	1

Table 7: Pair-wise matrix of alternatives based on robustness

	SVM	NB	DT	RF	LR
SVM	1	2.545	1.818	1.428	1.705
NB	0.393	1	1.522	1.524	1.698
DT	0.550	0.657	1	1.169	1.781
RF	0.700	0.656	0.856	1	1.753
LR	0.587	0.589	0.562	0.570	1

5.6. Perform a consistency check

Validity of priority factors is contingent upon Consistency Ratio (CR) values below 0.1, with CR values up to 0.1 deemed acceptable. To enhance the consistency of the pair-wise Table, iteration of the process is necessary if CR values deviate from the recommended range. Within the AHP, the evaluation of pair-wise matrix consistency is conducted through the assessment of Consistency Index (CI) and Consistency Ratio (CR). Eq. 13 is utilized to evaluate the uniformity of the pair-wise comparison matrix.

$$CI = \frac{\lambda_{max}-1}{n-1} \tag{13}$$

In this context, CI stands for the Consistency Index, λ_{max} represents the largest eigenvalue of the matrix, and \$n\$ refers to the matrix dimensions. Once the CI is found, the Consistency Ratio is determined using Eq. 14.

$$CR = \frac{CI}{RI} \tag{14}$$

where, CR denotes the consistency ratio, CI is employed for the consistency index, and RI refers to the random consistency index presented in Table 4, featuring fixed values. The weighted value (W) for each ML algorithm is derived by computing the average of the normalized values within the corresponding row, as illustrated in Tables 9-12. Consequently, the calculation of λ_{max} for each category is presented alongside the Normalized Tables 9-12.

Table 8: Pair-wise matrix of alternatives based on interpretability

	SVM	NB	DT	RF	LR
SVM	1	1.877	1.488	1.452	1.363
NB	0.533	1	1.663	1.775	1.616
DT	0.672	0.601	1	1.664	1.426
RF	0.689	0.563	0.601	1	1.302
LR	0.734	0.619	0.701	0.768	1

Table 9: Normalized matrix for pair-wise comparison of criteria

	Accuracy	Robustness	Interpretable	Priority weight
Accuracy	0.548	0.622	0.452	0.541
Robustness	0.212	0.241	0.349	0.267
Interpretable	0.241	0.137	0.199	0.192

Table 10: Normalized matrix for ML algorithms based on accuracy

	SVM	NB	DT	RF	LR	Priority weight
SVM	0.301	0.425	0.314	0.251	0.234	0.305
NB	0.120	0.170	0.241	0.204	0.198	0.186
DT	0.171	0.126	0.178	0.226	0.213	0.183
RF	0.248	0.172	0.162	0.206	0.230	0.204
LR	0.161	0.107	0.104	0.112	0.125	0.122

Table 11: Normalized matrix for ML algorithms based on robustness

	SVM	NB	DT	RF	LR	Priority weight
SVM	0.310	0.467	0.316	0.251	0.215	0.312
NB	0.122	0.184	0.264	0.268	0.214	0.210
DT	0.170	0.121	0.174	0.205	0.224	0.179
RF	0.217	0.120	0.149	0.176	0.221	0.177
LR	0.182	0.108	0.098	0.100	0.126	0.123

5.7. Calculating the local weight (LW) and global weight (GW)

The local weight of an Algorithm refers to the priority weight assigned to each Algorithm for each criterion. Consequently, at this stage, all the Algorithm priority weights are calculated and listed relative to each criterion. The local weight of each Algorithm for every criterion is multiplied by the weight of the corresponding criterion to obtain the global weight of each Algorithm. Both the local weight (LW) and global weight (GW) have been computed and are presented in Table 13.

5.8. Identify and create the overall priority ranking

In this step, we compile the ultimate selection of machine learning algorithms for text mining by evaluating their global weight. Algorithms with higher global weight values in all categories are deemed to be ranked more favorably. To determine the final priority, we sum up the global weights of each algorithm across all criteria.

In this survey, a comparison was conducted among the five most employed and top-performing machine learning algorithms, and their rankings were determined based on their global weight. A higher global weight indicates a higher level of significance. Among the five algorithms, SVM emerged as the most important, while LR was ranked as the least important. The results are presented in Table 14, corroborating the findings of our earlier SLR, which highlighted SVM as both the most used and best-performing algorithm.

6. Experimental analysis of machine learning models for emotion detection

To analyze the performance of machine learning models for emotion detection, we conducted multiple experiments to check the influence of text preprocessing, hyperparameter tuning, ensemble modeling techniques like stacking and boosting, and various vectorization methods such as Bag-of-Words, TF-IDF, and Word2Vec. For the above experiments, we selected the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset, which has been repeatedly used by different researchers working on text-based emotion classification. Fig. 8 shows the summary of the ISEAR dataset.

6.1. Comparative analysis of ML models before and after preprocessing

To analyze the impact of text preprocessing, a set of experiments was conducted. The experiments were performed on the ISEAR dataset using a suite of machine learning algorithms: K-Nearest Neighbors (KNN), Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and XGBoost (XGB). The initial expectation was that text preprocessing would enhance the performance of these models by filtering out noise and irrelevant information. However, the results indicated otherwise, with most models showing a decrease in performance post-preprocessing. Upon investigation, the cause of this decline was attributed to the removal of stop words. Common NLP libraries such as NLTK have a predefined list of stop words, which includes words that are often critical in understanding sentiment and emotion, like negations and contractions (e.g., "not," "don't," "didn't," "hasn't").

When these words are removed, the emotional context of the text can be altered, leading to poorer performance of models in detecting the intended sentiment as shown in Fig. 9. In contrast, when text processing was handled by an advanced library like SpaCy, the performance of the models improved as shown in Fig. 10. SpaCy's preprocessing appears to be more intelligent in handling stop words, by preserving words that are significant for understanding sentiment.

The pre-defined lists of stop words may need reassessment, especially for tasks like sentiment analysis and emotion detection, where every word can carry weight. NLP practitioners might need to customize stop word lists or develop algorithms that dynamically identify stop words based on the task at hand.

Table 12: Normalized matrix for ML algorithms based on interpretability

	SVM	NB	DT	RF	LR	Priority weight
SVM	0.276	0.403	0.273	0.218	0.203	0.274
NB	0.147	0.215	0.305	0.267	0.241	0.235
DT	0.185	0.129	0.183	0.250	0.213	0.192
RF	0.190	0.121	0.110	0.150	0.194	0.153
LR	0.202	0.133	0.129	0.115	0.149	0.146

Table 13: Summary of local and global weights of ML algorithms and their rankings

Criteria	Criteria weight	Algorithms	Local weights	Local ranking	Global weights	Final priority
Accuracy	0.5407	SVM	0.3051	1	0.1650	1
		NB	0.1865	3	0.1008	3
		DT	0.1829	4	0.0989	4
		RF	0.2036	2	0.1101	2
		LR	0.1219	5	0.0659	6
Robustness	0.2671	SVM	0.3116	1	0.0832	5
		NB	0.2103	2	0.0562	7
		DT	0.1789	3	0.0478	9
		RF	0.1765	4	0.0471	10
		LR	0.1227	5	0.0328	13
Interpretability	0.1922	SVM	0.2745	1	0.0528	8
		NB	0.2348	2	0.0451	11
		DT	0.1920	3	0.0369	12
		RF	0.1531	4	0.0294	14
		LR	0.1456	5	0.0280	15

Table 14: Final priority of ML algorithms

S. No.	ML Algorithm	Summing global weights	Priority
1	SVM	0.300983803	1
2	NB	0.2021202	2
3	RF	0.186639047	3
4	DT	0.18359513	4
5	LR	0.12666182	5

6.2. Comparative analysis of ML models before and after hyperparameter tuning

This experiment investigates the impact of hyperparameter tuning on the performance of the same seven machine learning models selected for the experiment. Two scenarios were considered: before and after hyperparameter tuning. Tuning involves adjusting model-specific parameters to maximize predictive accuracy. By Grid Search, the best hyperparameter values of each model for our selected dataset are shown in Table 14.

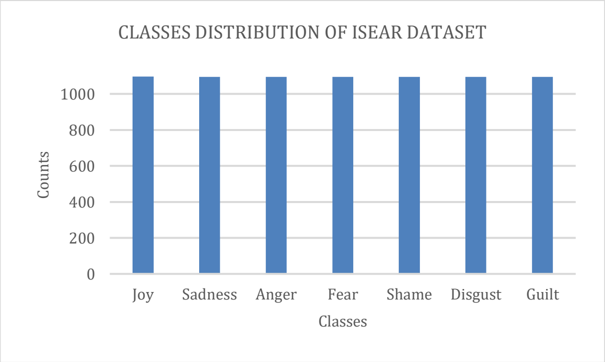


Fig. 8: Class distribution of the ISEAR dataset

The accuracy scores for each model before and after hyperparameter tuning are shown in Fig. 11. The results demonstrate notable improvements in accuracy for several models, highlighting the importance of optimizing model configurations. SVM, RF, and LR showed significant accuracy gains after tuning. KNN exhibited a considerable accuracy increase from 43% to 49% while NB and LR maintained consistent accuracy before and after hyperparameter tuning.

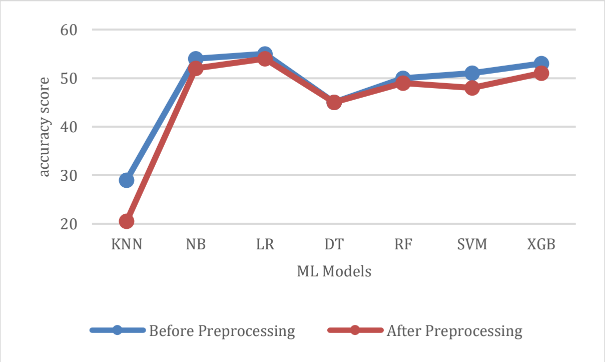


Fig. 9: Comparison of ML model before and after reprocessing (with NLTK)

Hyperparameter tuning allows models to better adapt to the specifics of the dataset, improving their generalization capabilities. Algorithms, like KNN and Decision Trees, are particularly sensitive to hyperparameter settings, leading to significant performance gains post-tuning. Algorithms like Naive Bayes and Logistic Regression, which have fewer hyperparameters, show less improvement.

6.3. Comparative analysis of vectorization methods

The purpose of this experiment was to investigate the impact of different vectorization techniques on the performance of the selected machine learning models. The selected vectorization techniques include Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and Word2Vec. In using Word2Vec, the pre-trained model obtained from extensive corpora of Google News is accessed through the 'gensim' library. Fig. 12 shows the accuracy comparison of the selected model for each vectorization technique. TF-IDF is a versatile option for different models, utilizing term importance and inverse document frequency. Bag-of-Words (BoW), known for its simplicity, is effective in models such as Naive Bayes (NB) and Logistic Regression (LR).

The outcome for Word2Vec was very low in most cases, which was contrary to our expectations. But this pattern of low accuracies with state-of-the-art vectorization techniques like GloVe, FastText, and Word2Vec on the ISEAR dataset is also reported by other studies (Saravia et al., 2018). TF-IDF generally performs well because it accounts for the importance of terms across the corpus, providing a balanced feature set. BoW can be more effective with algorithms that do not require weighted features, such as Decision Trees. Word2Vec embeddings are dense and capture semantic relationships, which benefit algorithms like SVM.

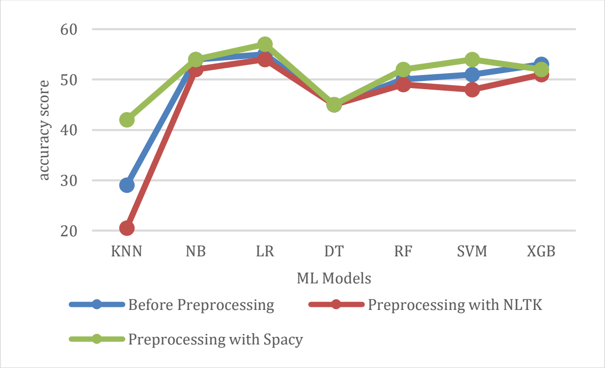


Fig. 10: Comparison of ML model before and after preprocessing, i.e., NLTK and SpaCy

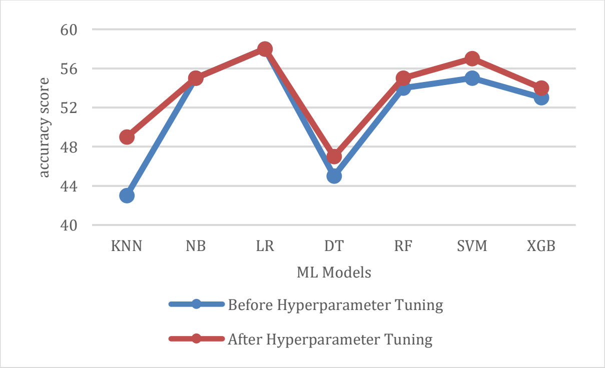


Fig. 11: Comparison of ML model with and without hyperparameter tuning

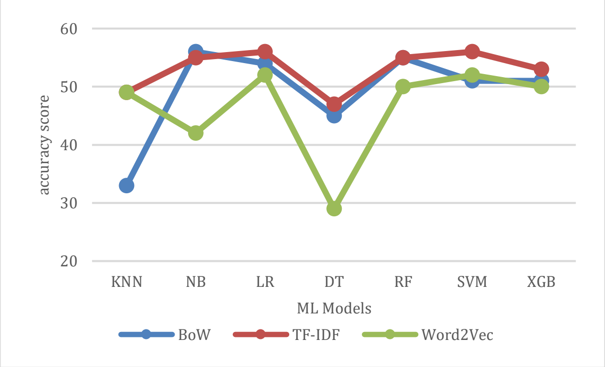


Fig. 12: Comparison of vectorization techniques, i.e., BoW, TF-IDF, and Word2Vec

6.4. Stacking multiple ML models for performance improvement

Stacking involves training multiple models to forecast the same target variable and then employing a meta-model to generate final predictions using outputs of the individual models. In other words, instead of using voting (as in bagging) or averaging (as in boosting), stacking leverages multiple models and combines their predictions using another model. In this experiment, multi-level stacking is used, which is more complex than simple stacking. In our case, we have three base models, two intermediate models, and finally one meta model; the architecture of multi-level stacking is shown in Fig. 13.

First, the base models are trained independently on the training data. The predictions from the base models serve as features for the intermediate-level models. The intermediate-level models make their predictions, which, along with the original features, serve as inputs for the final meta-model at the third level. The meta-model learns to combine the predictions of the intermediate-level models to make a final prediction. Now the trained meta-model can be used to make predictions on unseen data. Based on the above experiments, we only consider NB, LR, RF, SVM, and XGB and ignore KNN and DT due to their consistently low performance. With these five models, we have performed many experiments with different combinations, and the top 4 best-performing stacking combinations are shown in Table 13. It can be clearly seen that multi-level stacking outperforms all base models. The ability to harness the strengths of multiple models, reduce individual errors, and adapt to complex and nuanced relationships in text data makes it more effective than individual models.

7. Evaluation of transfer learning models for emotion detection

The adoption of transformer-based models in this study was driven by their ability to address the limitations of conventional machine learning models in emotion detection. By leveraging the powerful contextualization capabilities of Bidirectional Encoder Representations from Transformers

(BERT), Robustly Optimized BERT Pretraining Approach (RoBERTa), Generalized Autoregressive Pretraining for Language Understanding (XLNet), Distilled Version of BERT (DistilBERT), and Decoding-enhanced BERT with Disentangled Attention (DeBERTa), this research aims to enhance the accuracy and robustness of emotion detection systems. These models offer significant advantages in capturing the subtleties and dynamics of human emotions in text, making them indispensable tools for advancing emotion detection methodologies.

7.1. Experimental setup

The Categorical Cross-Entropy loss function was applied to measure the difference between predicted outputs and actual labels. This method is widely used for multi-class classification tasks. The models were fine-tuned using the Adam optimizer, chosen for its ability to handle sparse gradients and provide adaptive learning rates effectively (Chan et al., 2023). In addition, the early stopping technique was implemented during training. This approach automatically stopped the training process when the model's performance on the validation set no longer improved, helping to reduce overfitting and ensure better generalization of the models to unseen data.

7.2. Results and discussion

Experimental outcomes reveal that the DeBERTa model attains an average accuracy of 76.5%, thus outperforming RoBERTa (74.31%), XLNet (72.99%), BERT (70.09%), and DistilBERT (66.93%). The training loss and test accuracy graphs of DeBERTa, RoBERTa, XLNet, BERT, and DistilBERT are shown in Figs. 14-18, respectively. DeBERTa has the advantage of the disentangled attention mechanism, which allows understanding the semantic dependencies and the positions to perform various emotion recognition tasks (Xian et al., 2023). The better performance of the model, including more effective identification of difficult emotions, such as guilt and shame (Assiri et al., 2024).

8. Limitations and challenges

Emotion recognition from text is a complex and challenging task due to several inherent limitations and challenges, particularly when leveraging ML models. In this section, we critically analyze the limitations and challenges of using conventional ML algorithms alongside advanced models, transfer learning models. We also provide actionable recommendations to address these challenges.

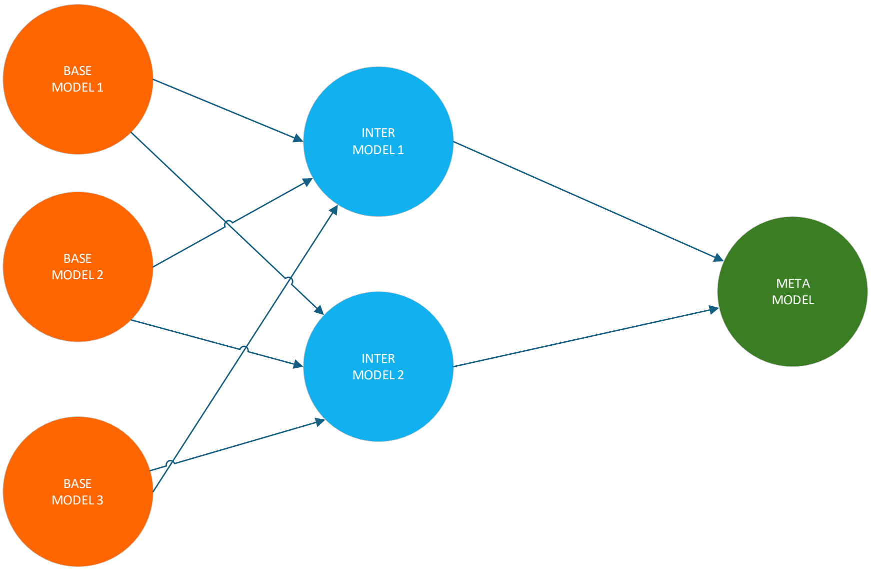


Fig. 13: Architecture of multi-level stacking

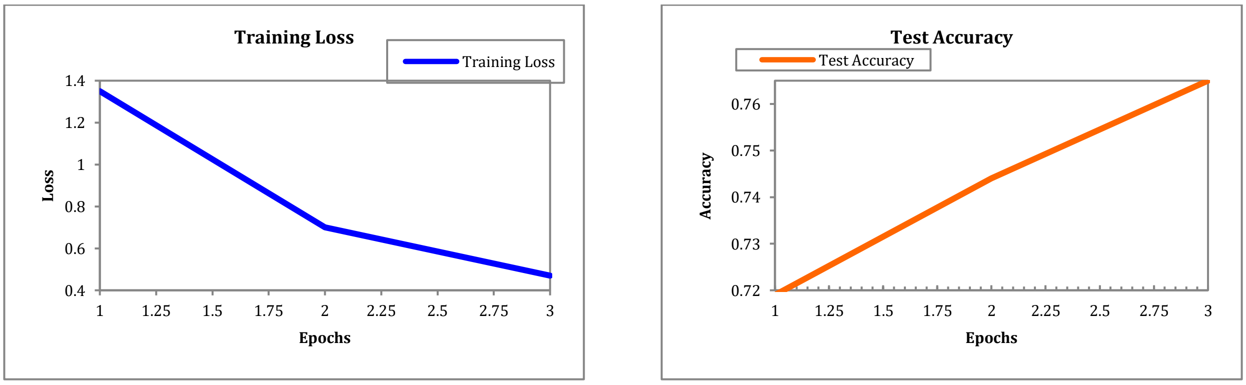


Fig. 14: DeBERTa: Training loss and test accuracy graph

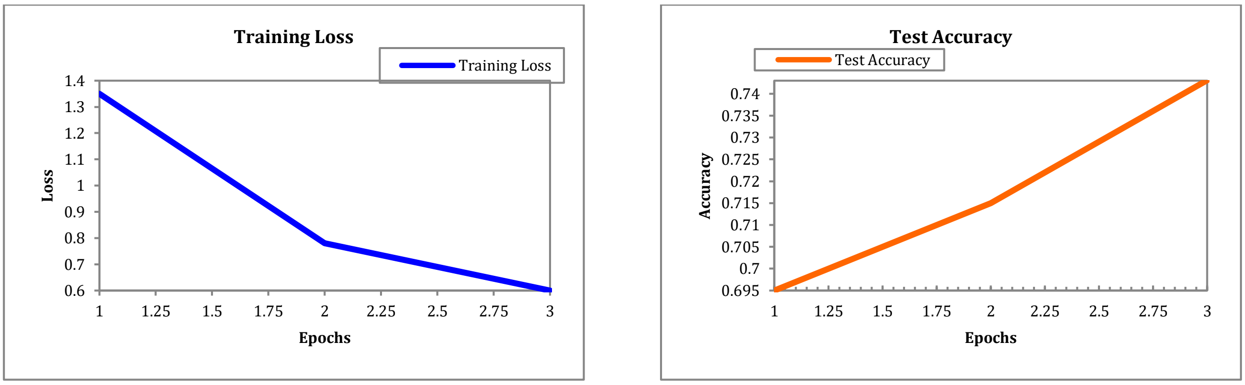


Fig. 15: RoBERTa: Training loss and test accuracy graph

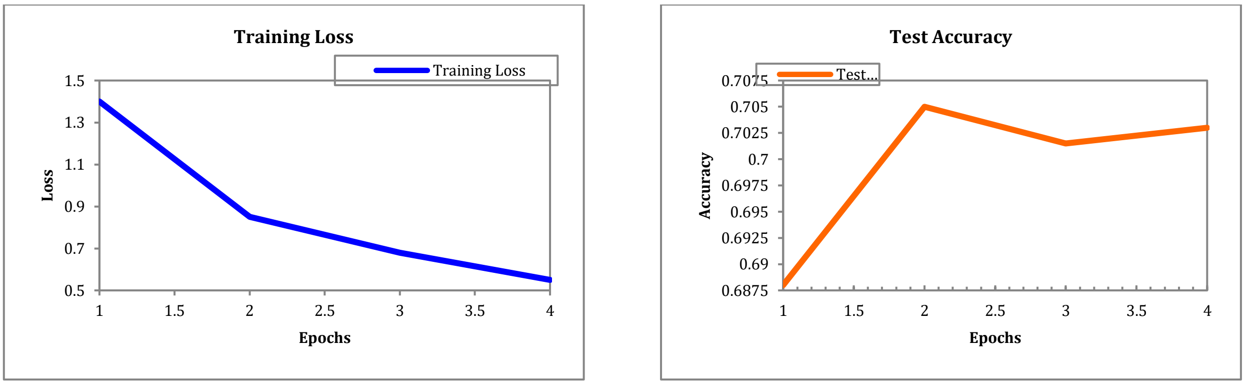


Fig. 16: XLNet: Training loss and test accuracy graph

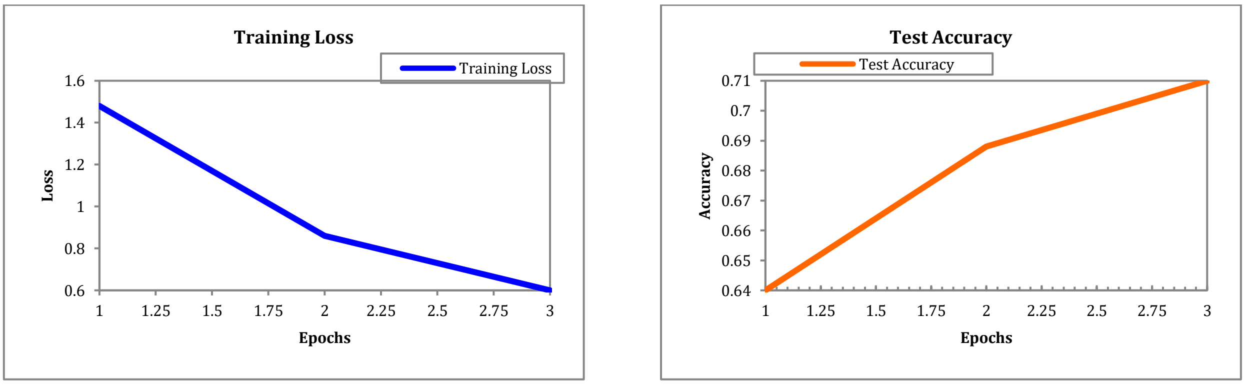


Fig. 17: BERT: Training loss and test accuracy graph

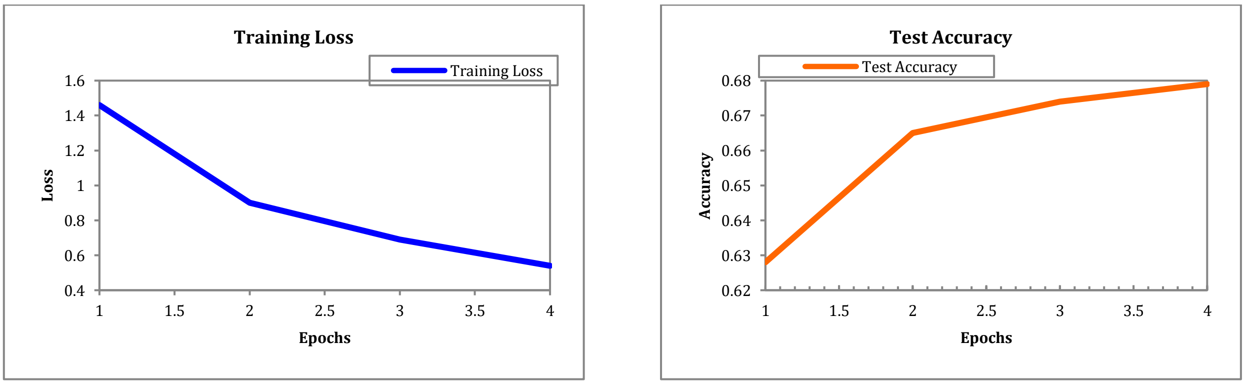


Fig. 18: DistilBERT: Training loss and test accuracy graph

8.1. Data scarcity and imbalance

One of the most critical challenges in emotion detection tasks is the lack of labeled emotional data. Emotion detection datasets are often limited in size and may not cover a diverse range of emotions and contexts. Moreover, datasets often suffer from class imbalance, where certain emotions are overrepresented, while others are underrepresented. This imbalance can lead to biased models that perform well on dominant emotions but poorly on less-represented ones. To address data scarcity and imbalance, it is crucial to augment existing datasets using data augmentation techniques such as back-translation or synthetic data generation. Additionally, employing techniques like oversampling, undersampling, or using class weights during model training can mitigate the effects of class imbalance. Researchers can also explore transfer learning to leverage pre-trained models on large corpora and fine-tune them on smaller emotion-specific datasets.

8.2. Contextual understanding

Text-based emotion detection heavily depends on the ability to understand the context of the text. Emotions can often be expressed indirectly or implicitly, and words or phrases may have different emotional connotations depending on the surrounding context.

Advanced transformer-based models like BERT, RoBERTa, and DeBERTa, which use contextual embeddings, can help improve emotion detection by capturing the nuances of words in different contexts. Fine-tuning these models with domain-specific data can further enhance the ability to detect emotions in varied contexts.

8.3. Generalization and overfitting

Many ML-based emotion detection models, especially those trained on small or domain-specific datasets, suffer from overfitting. This happens when a model memorizes the training data rather than generalizing to unseen data. Even advanced models like BERT and RoBERTa, though more robust, can still overfit if they are fine-tuned excessively on small datasets. To prevent overfitting, regularization techniques such as dropout, L2 regularization, or early stopping should be employed during model training.

8.4. Interpretability and transparency

One significant challenge with complex deep learning models like BERT, RoBERTa, and XLNet is the lack of interpretability. These models are often considered "black boxes," making it difficult to understand how specific emotions are being predicted. This can be particularly problematic in real-world applications where explainability is crucial. To enhance model interpretability, techniques like attention visualization, Shapley Additive Explanations (SHAP), and Local Interpretable Model-Agnostic Explanations (LIME) can be applied.

8.5. Cross-cultural and multilingual issues

Emotion expression can vary significantly across cultures and languages. A model trained on data from one language or culture might not perform well when applied to others. To address this issue, researchers should focus on developing multilingual and cross-cultural emotion recognition models. Pre-trained models like Multilingual BERT (M-BERT) or Cross-lingual RoBERTa (XLM-R) on large, multilingual datasets can help build models that generalize better across languages.

9. Conclusion

The processing of the ever-growing volume of data manually, particularly in textual form, poses a challenge for individuals. Text mining comprises a range of methods employed to extract valuable information and discern complex patterns from textual data. Our study unfolded through three main aspects: an in-depth literature review, an AHP survey from field experts, and a set of practical experiments. In our literature review, the prevalence of SVM as the foremost choice in 72% of studies underscores its efficacy in handling non-linear boundaries and managing class imbalance. Notably, Naïve Bayes (NB) follows closely at 56%, indicating its continued relevance in the field. The consistent outperformance of SVM reaffirms its robustness, while customized datasets find favor in 80% of studies, with the Ekman model featuring six emotion classes emerging as a popular choice (Nandwani and Verma, 2021). Our analysis highlights that datasets containing four to eight emotion classes yield optimal accuracy, emphasizing the importance of tailored dataset curation. The insights garnered from the AHP survey, which engaged industry experts, align with the literature review, as SVM is recommended as the optimal choice, reaffirming its prominence. Naïve Bayes and Random Forest (RF) emerge as the second and third choices based on criteria such as accuracy, robustness, and interpretability. This convergence of expert opinions further reinforces the credibility of SVM in real-world applications. Our practical experiments provided a hands-on validation of the literature and expert recommendations. Support Vector Machine, Logistic Regression (LR), and Naïve Bayes consistently demonstrated robust performance. Furthermore, our exploration into ensemble techniques revealed multi-level stacking as the most effective classifier, surpassing individual models, particularly outperforming the leading SVM by 3%.

We have also performed experiments with transformer-based models, including BERT, RoBERTa, XLNet, DistilBERT, and DeBERTa, to further enhance the accuracy and robustness of emotion detection systems. The findings reveal that the DeBERTa model outperformed other models. These results highlight the effectiveness of transfer learning models in capturing the nuances of human emotions in text and demonstrate their potential for improving emotion detection systems.

List of abbreviations

AHP	Analytical hierarchy process
SLR	Systematic literature review
ML	Machine learning
SVM	Support vector machine

NB	Naive bayes
RF	Random forest
DT	Decision tree
KNN	K-nearest neighbor
LR	Logistic regression
XGB	XGBoost
ED	Emotion detection
TM	Text mining
NLP	Natural language processing
BoW	Bag-of-words
TF-IDF	Term frequency-inverse document frequency
VSM	Vector space model
PMI	Pointwise mutual information
POS	Part-of-speech
CV	Cross-validation
CI	Consistency index
CR	Consistency ratio
RI	Random consistency index
LW	Local weight
GW	Global weight
ISEAR	International survey on emotion antecedents and reactions
BERT	Bidirectional encoder representations from transformers
RoBERTa	Robustly optimized BERT pretraining approach
XLNet	Generalized autoregressive pretraining for language understanding
DistilBERT	Distilled version of BERT
DeBERTa	Decoding-enhanced BERT with disentangled attention
SHAP	Shapley additive explanations
LIME	Local interpretable model-agnostic explanations
M-BERT	Multilingual BERT
XLM-R	Cross-lingual RoBERTa

Acknowledgment

The authors extend their appreciation to the Arab Open University for funding this work through the AOU research fund No. (AOUKSA524008).

Compliance with ethical standards

Ethical considerations

The authors confirm that participation in the expert survey was voluntary. Informed consent was obtained from all participants prior to data collection. No personal or sensitive information was collected, and all responses were anonymized to ensure confidentiality.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abdolvand N, Albadvi A, and Aghdasi M (2015). Performance management using a value-based customer-centered model. International Journal of Production Research, 53(18): 5472–5483. https://doi.org/10.1080/00207543.2015.1026613

Abdullah M, AlMasawa M, Makki I, Alsolmi M, and Mahrous S (2020). Emotions extraction from Arabic tweets. International Journal of Computers and Applications, 42(7): 661–675. https://doi.org/10.1080/1206212X.2018.1482395

Abrar MF, Khan MS, Khan I, Ali G, and Shah S (2023b). Digital information credibility: Towards a set of guidelines for quality assessment of grey literature in multivocal literature review. Applied Sciences, 13(7): 4483. https://doi.org/10.3390/app13074483

Abrar MF, Khan MS, Khan I, ElAffendi M, and Ahmad S (2023a). Towards fake news detection: A multivocal literature review of credibility factors in online news stories and analysis using analytical hierarchical process. Electronics, 12(15): 3280. https://doi.org/10.3390/electronics12153280

Acheampong FA, Wenyu C, and Nunoo-Mensah H (2020). Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports, 2(7): e12189.

Almahdawi A and Teahan WJ (2018). Automatically recognizing emotions in text using prediction by partial matching (PPM) text compression method. In the New Trends in Information and Communications Technology Applications: 3rd International Conference, Springer International Publishing, Baghdad, Iraq: 269-283. https://doi.org/10.1007/978-3-030-01653-1_17

Alotaibi FM (2019). Classifying text-based emotions using logistic regression. VAWKUM Transactions on Computer Sciences, 7(1): 31–37. https://doi.org/10.21015/vtcs.v16i2.551

Alswaidan N and Menai MEB (2020). A survey of state-of-the-art approaches for emotion recognition in text. Knowledge and Information Systems, 62(8): 2937-2987. https://doi.org/10.1007/s10115-020-01449-0

Angel Deborah S, Rajalakshmi S, Milton Rajendram S, and Mirnalinee TT (2020). Contextual emotion detection in text using ensemble learning. In: Hemanth DJ, Kumar VDA, Malathi S, Castillo O, and Patrut B (Eds.), Emerging trends in computing and expert technology: 1179–1186. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-030-32150-5_121

Assiri A, Gumaei A, Mehmood F, Abbas T, and Ullah S (2024). DeBERTa-GRU: Sentiment analysis for large language model. Computers, Materials and Continua, 79(3): 4219-4236. https://doi.org/10.32604/cmc.2024.050781

Balakrishnan V and Kaur W (2019). String-based multinomial Naïve Bayes for emotion detection among Facebook diabetes community. Procedia Computer Science, 159: 30–37. https://doi.org/10.1016/j.procs.2019.09.157

Balakrishnan V, Lok PY, and Abdul Rahim H (2021). A semi-supervised approach in detecting sentiment and emotion based on digital payment reviews. The Journal of Supercomputing, 77(4): 3795–3810. https://doi.org/10.1007/s11227-020-03412-w

Carrera-Rivera A, Ochoa W, Larrinaga F, and Lasa G (2022). How-to conduct a systematic literature review: A quick guide for computer science research. MethodsX, 9: 101895. https://doi.org/10.1016/j.mex.2022.101895 **PMid:36405369 PMCid:PMC9672331**

Chaffar S and Inkpen D (2011). Using a heterogeneous dataset for emotion analysis in text. In the Advances in Artificial Intelligence: 24th Canadian Conference on Artificial Intelligence, Canadian AI 2011, Springer Berlin Heidelberg, St. John's, Canada: 62-67. https://doi.org/10.1007/978-3-642-21043-3_8

Chakriswaran P, Vincent DR, Srinivasan K, Sharma V, Chang CY, and Reina DG (2019). Emotion AI-driven sentiment analysis: A survey, future research directions, and open issues. Applied Sciences, 9(24): 5462. https://doi.org/10.3390/app9245462

Chan YL, Bea KT, Leow SMH, Phoong SW, and Cheng WK (2023). State of the art: A review of sentiment analysis based on sequential transfer learning. Artificial Intelligence Review,

56(1): 749–780.
https://doi.org/10.1007/s10462-022-10183-8

Chowanda A, Sutoyo R, and Tanachutiwat S (2021). Exploring text-based emotions recognition machine learning techniques on social media conversation. Procedia Computer Science, 179: 821-828. https://doi.org/10.1016/j.procs.2021.01.099

Esmin AA, De Oliveira Jr RL, and Matwin S (2012). Hierarchical classification approach to emotion recognition in Twitter. In the 11th International Conference on Machine Learning and Applications, IEEE, Boca Raton, USA, 2: 381-385. https://doi.org/10.1109/ICMLA.2012.195

Ghanbari-Adivi F and Mosleh M (2019). Text emotion detection in social networks using a novel ensemble classifier based on Parzen tree estimator (TPE). Neural Computing and Applications, 31(12): 8971-8983. https://doi.org/10.1007/s00521-019-04230-9

Gohil L and Patel D (2019). Multilabel classification for emotion analysis of multilingual tweets. International Journal of Innovative Technology and Exploring Engineering, 9(1): 4453–4457. https://doi.org/10.35940/ijitee.A5320.119119

Gunarathne SR, De Silva J, Ekanayake EP, Samaradiwakara I, Haddela PS, and Fernando PA (2013). Intellemo: A mobile instant messaging application with intelligent emotion identification. In the IEEE 8th International Conference on Industrial and Information Systems, IEEE, Peradeniya, Sri Lanka: 627-632. https://doi.org/10.1109/ICIInfS.2013.6732057

Halczak P (2023). Dictionary representation of the semantics of adjectives signifying emotions. International Journal of Lexicography, 36(4): 424–446. https://doi.org/10.1093/ijl/ecad016

Halim Z, Waqar M, and Tahir M (2020). A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email. Knowledge-Based Systems, 208: 106443. https://doi.org/10.1016/j.knosys.2020.106443

Hussein A, Al Kafri M, Abonamah AA, and Tariq MU (2020). Mood detection based on Arabic text documents using machine learning methods. International Journal, 9(4): 4424-4436. https://doi.org/10.30534/ijatcse/2020/36942020

Jain VK, Kumar S, and Fernandes SL (2017). Extraction of emotions from multilingual text using intelligent text processing and computational linguistics. Journal of Computational Science, 21: 316-326. https://doi.org/10.1016/j.jocs.2017.01.010

Kabra G, Ramesh A, and Arshinder K (2015). Identification and prioritization of coordination barriers in humanitarian supply chain management. International Journal of Disaster Risk Reduction, 13: 128-138. https://doi.org/10.1016/j.ijdrr.2015.01.011

Kalcheva N, Karova M, and Penev I (2020). Comparison of the accuracy of SVM kemel functions in text classification. In the International Conference on Biomedical Innovations and Applications (BIA), IEEE, Varna, Bulgaria: 141-145. https://doi.org/10.1109/BIA50171.2020.9244278

Kang X, Ren F, and Wu Y (2017). Exploring latent semantic information for textual emotion recognition in blog articles. IEEE/CAA Journal of Automatica Sinica, 5(1): 204–216. https://doi.org/10.1109/JAS.2017.7510421

Kaur W, Balakrishnan V, and Singh B (2020). Improving teaching and learning experience in engineering education using sentiment analysis techniques. IOP Conference Series: Materials Science and Engineering, 834(1): 12026. https://doi.org/10.1088/1757-899X/834/1/012026

Krenicky T, Hrebenyk L, and Chernobrovchenko V (2022). Application of concepts of the analytic hierarchy process in decision-making. Management Systems in Production Engineering, 4(30): 304-310. https://doi.org/10.2478/mspe-2022-0039

Kumar A, Nadeem M, and Shameem M (2023). Systematic literature review of metrics for measuring devops success. AIP Conference Proceedings, 2724(1): 030005. https://doi.org/10.1063/5.0128883

Liu L and Qi J (2018). Research on discrete emotion classification of Chinese online product reviews based on OCC model. In the IEEE 3rd International Conference on Data Science in Cyberspace, IEEE, Guangzhou, China: 371-378. https://doi.org/10.1109/DSC.2018.00060

Mahajan R and Zaveri M (2021). Harnessing emotive features for emotion recognition from text. International Journal of Advanced Computer Science and Applications, 12(7): 166-175. https://doi.org/10.14569/IJACSA.2021.0120719

Majeed A, Mujtaba H, and Beg MO (2020). Emotion detection in Roman Urdu text using machine learning. In the Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering, ACM, Virtual Event, Australia: 125-130. https://doi.org/10.1145/3417113.3423375

Mondal A and Gokhale SS (2020). Mining emotions on Plutchik's wheel. In the 7th International Conference on Social Networks Analysis, Management and Security, IEEE, Paris, France: 1-6. https://doi.org/10.1109/SNAMS52053.2020.9336534

Munezero M, Montero CS, Sutinen E, and Pajunen J (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. IEEE Transactions on Affective Computing, 5(2): 101–111. https://doi.org/10.1109/TAFFC.2014.2317187

Murthy AR and Kumar KMA (2021). A review of different approaches for detecting emotion from text. IOP Conference Series: Materials Science and Engineering, 1110(1): 12009. https://doi.org/10.1088/1757-899X/1110/1/012009

Nandwani P and Verma R (2021). A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining, 11(1): 81. https://doi.org/10.1007/s13278-021-00776-6 **PMid:34484462 PMCid:PMC8402961**

Nguyen KP-Q and Van Nguyen K (2020). Exploiting Vietnamese social media characteristics for textual emotion recognition in Vietnamese. In the International Conference on Asian Language Processing (IALP): 276–281. https://doi.org/10.1109/IALP51396.2020.9310495

Pang J, Rao Y, Xie H, Wang X, Wang FL, Wong T-L, and Li Q (2019). Fast supervised topic models for short text emotion detection. IEEE Transactions on Cybernetics, 51(2): 815–828. https://doi.org/10.1109/TCYB.2019.2940520 **PMid:31567111**

Parvin T and Hoque MM (2021). An ensemble technique to classify multi-class textual emotion. Procedia Computer Science, 193: 72–81. https://doi.org/10.1016/j.procs.2021.10.008

Patacsil FF (2020). Emotion recognition from blog comments based automatically generated datasets and ensemble models. International Journal, 9(4): 5979-5986. https://doi.org/10.30534/ijatcse/2020/264942020

Patil T and Patil S (2013). Automatic generation of emotions for social networking websites using text mining. In the 4th International Conference on Computing, Communications and Networking Technologies, IEEE, Tiruchengode, India: 1-6. https://doi.org/10.1109/ICCCNT.2013.6726704

Paul J and Barari M (2022). Meta-analysis and traditional systematic literature reviews—What, why, when, where, and how? Psychology and Marketing, 39(6): 1099–1115. https://doi.org/10.1002/mar.21657

Plaza-del-Arco FM, Martín-Valdivia MT, Urena-Lopez LA, and Mitkov R (2020). Improved emotion recognition in Spanish social media through incorporation of lexical knowledge. Future Generation Computer Systems, 110: 1000-1008. https://doi.org/10.1016/j.future.2019.09.034

Povoda L, Arora A, Singh S, Burget R, and Dutta MK (2015). Emotion recognition from helpdesk messages. In the 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, IEEE, Brno, Czech Republic: 310-313. https://doi.org/10.1109/ICUMT.2015.7382448

Povoda L, Burget R, Masek J, Uher V, and Dutta MK (2016). Optimization methods in emotion recognition system. Radioengineering, 25(3): 565–572. https://doi.org/10.13164/re.2016.0565

Putra OV, Wasmanson FM, Harmini T, and Utama SN (2020). Sundanese Twitter dataset for emotion classification. In the International Conference on Computer Engineering, Network, and Intelligent Multimedia, IEEE, Surabaya, Indonesia: 391-395. https://doi.org/10.1109/CENIM51130.2020.9297929

Ray A, Bala PK, and Jain R (2021). Utilizing emotion scores for improving classifier performance for predicting customer's intended ratings from social media posts. Benchmarking: An International Journal, 28(2): 438–464. https://doi.org/10.1108/BIJ-01-2020-0004

Saad MM, Jamil N, and Hamzah R (2018). Evaluation of support vector machine and decision tree for emotion recognition of Malay folklores. Bulletin of Electrical Engineering and Informatics, 7(3): 479–486. https://doi.org/10.11591/eei.v7i3.1279

Sailunaz K and Alhajj R (2019). Emotion and sentiment analysis from Twitter text. Journal of Computational Science, 36: 101003. https://doi.org/10.1016/j.jocs.2019.05.009

Saputri MS, Mahendra R, and Adriani M (2018). Emotion classification on Indonesian Twitter dataset. In the International Conference on Asian Language Processing, IEEE, Bandung, Indonesia: 90-95. https://doi.org/10.1109/IALP.2018.8629262

Sarakit P, Theeramunkong T, Haruechaiyasak C, and Okumura M (2015). Classifying emotion in Thai YouTube comments. In the 6th International Conference of Information and Communication Technology for Embedded Systems, IEEE, Hua Hin, Thailand: 1-5. https://doi.org/10.1109/ICTEmSys.2015.7110808

Saravia E, Liu HCT, Huang YH, Wu J, and Chen YS (2018). CARER: Contextualized affect representations for emotion recognition. In the Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium: 3687-3697. https://doi.org/10.18653/v1/D18-1404

Shameem M, Kumar RR, Kumar C, Chandra B, and Khan AA (2018). Prioritizing challenges of agile process in distributed software development environment using analytic hierarchy process. Journal of Software: Evolution and Process, 30(11): e1979. https://doi.org/10.1002/smr.1979

Sintsova V, Musat C, and Pu P (2014). Semi-supervised method for multi-category emotion recognition in tweets. In the IEEE International Conference on Data Mining Workshop, IEEE, Shenzhen, China: 393-402. https://doi.org/10.1109/ICDMW.2014.146

Sreeja PS and Mahalakshmi GS (2019). Emotion recognition in poetry using ensemble of classifiers. In the Next Generation Computing Technologies on Computational Intelligence: 4th International Conference, Springer Singapore, Dehradun, India: 77-91. https://doi.org/10.1007/978-981-15-1718-1_7

Suhasini M and Srinivasu B (2020). Emotion detection framework for Twitter data using supervised classifiers. In the Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19, Springer Nature, Singapore, Singapore: 565-576. https://doi.org/10.1007/978-981-15-1097-7_47

Tian F, Gao P, Li L, Zhang W, Liang H, Qian Y, and Zhao R (2014). Recognizing and regulating e-learners' emotions based on interactive Chinese texts in e-learning systems. Knowledge-Based Systems, 55: 148-164. https://doi.org/10.1016/j.knosys.2013.10.019

Tuhin RA, Paul BK, Nawrine F, Akter M, and Das AK (2019). An automated system of sentiment analysis from Bangla text using supervised learning techniques. In the IEEE 4th International Conference on Computer and Communication Systems, IEEE, Singapore, Singapore: 360-364. https://doi.org/10.1109/CCOMS.2019.8821658

Xian G, Guo Q, Zhao Z, Luo Y, and Mei H (2023). Short text classification model based on DeBERTa-DPCNN. In the 4th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, IEEE, Hangzhou, China: 56-59. https://doi.org/10.1109/ICBAIE59714.2023.10281320

Yuan Z and Purver M (2015). Predicting emotion labels for Chinese microblog texts. In: Gaber M, Cocea M, Wiratunga N, Goker A (Eds.), Advances in social media analysis: 129-149. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-18458-6_7

Zhang F, Xu H, Wang J, Sun X, and Deng J (2016). Grasp the implicit features: Hierarchical emotion classification based on topic model and SVM. In the International Joint Conference on Neural Networks, IEEE, Vancouver, Canada: 3592-3599. https://doi.org/10.1109/IJCNN.2016.7727661