IJAIA

Volume 1 | Number 2 | December 2025 | Pages 1-15

International Journal of Artificial Intelligence Applications

ISSN (online): 3105-3572
Homepage: ijaia.com

Performance Evaluation of Classical Machine Learning Models for Emotion Classification

Motaz Zghoul[1*], Amneh Shaban[2]

[1] Department of Artificial Intelligence, Faculty of Science and Information Technology,
Al-Zaytoonah University of Jordan, Amman 11733, Jordan, Email:Motazzghoul98@gmail.com
[2] Software Department, Faculty of Information Technology, Applied Science Private University,
Amman 11931, Jordan, Email: amnehshaban@hotmail.com

Received: 08, 2025 *Revised: 09, 2025* *Accepted: 10, 2025* *Available online: 12, 2025*

ABSTRACT — Emotion detection in textual data represents a critical challenge in natural language processing with applications in mental health monitoring, customer sentiment analysis, and human-computer interaction. This study investigates three classical machine learning algorithms for multi-class emotion classification across eleven emotional categories using a balanced dataset of approximately 106,000 annotated sentences. The research employs Term Frequency-Inverse Document Frequency vectorization with trigram support and 3,000-dimensional feature space. Logistic Regression, Random Forest, and Naive Bayes classifiers were evaluated using comprehensive metrics including accuracy, precision, recall, F1-score, and five-fold cross-validation. Results demonstrate that Logistic Regression achieved superior performance with 79.90% accuracy, 81.18% precision, and 80.27% F1-score, substantially exceeding Random Forest at 75.32% and Naive Bayes at 69.01%. Cross-validation analysis revealed remarkable stability with standard deviations below 0.5%, confirming robust generalization. Per-class analysis identified enthusiasm, love, and neutral as most reliably detected emotions exceeding 83% accuracy, while empty and sadness presented greater challenges. The findings validate that classical machine learning approaches with proper feature engineering achieve competitive performance for fine-grained emotion detection while offering advantages in computational efficiency, interpretability, and deployment simplicity.

Keywords — Emotion detection, text classification, machine learning, Logistic Regression, Random Forest, Naive Bayes, natural language processing, sentiment analysis, TF-IDF vectorization.

1. INTRODUCTION

Emotion detection in textual data represents a fundamental challenge in natural language processing with profound implications for human-computer interaction, mental health monitoring, customer sentiment analysis, and social media analytics. The ability to automatically identify and classify emotional states expressed through written language enables machines to better understand human communication, facilitating more empathetic and context-aware artificial intelligence systems. As digital communication continues to dominate human interaction through social media platforms, messaging applications, and online forums, the volume of emotion-laden text data has grown exponentially, creating both opportunities and challenges for automated emotion recognition systems [1].

Traditional approaches to emotion analysis have focused primarily on polarity-based sentiment classification, distinguishing between positive, negative, and neutral sentiments. However, this coarse-grained categorization fails to capture the rich emotional landscape inherent in human communication, where discrete emotions such as joy, sadness, anger, fear, surprise, and disgust manifest through distinct linguistic patterns and contextual cues [2]. The

[*] Corresponding author

DOI: https://doi.org/

transition from simple sentiment analysis to fine-grained emotion detection requires sophisticated natural language processing techniques capable of distinguishing subtle differences in emotional expression while maintaining robust performance across diverse textual contexts and communication styles [3].

Machine learning approaches have emerged as the predominant methodology for emotion detection tasks, leveraging large annotated corpora to learn statistical patterns that characterize different emotional states. Classical machine learning algorithms, including Support Vector Machines, Random Forests, and Naive Bayes classifiers, have demonstrated considerable success when combined with carefully engineered features such as bag-of-words representations, n-grams, and lexicon-based emotional indicators [4]. These traditional methods benefit from interpretability and computational efficiency, making them particularly suitable for applications requiring transparent decision-making processes or operating under resource constraints [5].

The advent of deep learning has revolutionized emotion detection capabilities through the development of neural architectures that automatically learn hierarchical feature representations from raw text data. Recurrent Neural Networks, particularly Long Short-Term Memory networks, have shown exceptional performance in capturing sequential dependencies and contextual information crucial for understanding emotional content [6]. More recently, transformer-based models such as BERT and its variants have achieved state-of-the-art performance across numerous natural language processing tasks, including emotion classification, by leveraging pre-trained language representations that encode rich semantic and syntactic knowledge [7].

Despite significant advances in deep learning methodologies, classical machine learning approaches remain highly relevant for emotion detection applications, particularly in scenarios involving limited computational resources, smaller datasets, or requirements for model interpretability. Traditional algorithms often provide competitive performance when combined with sophisticated feature engineering and appropriate preprocessing techniques, while offering advantages in training efficiency and transparency [8]. Furthermore, classical methods serve as essential baselines for evaluating more complex architectures and provide valuable insights into the fundamental characteristics of emotion-discriminative features in text data [9].

The present study investigates the application of three classical machine learning algorithms for multi-class emotion detection across eleven distinct emotional categories. The research employs a comprehensive emotion dataset comprising approximately 106,000 annotated sentences, specifically curated to support transformer-based architectures while maintaining balance across emotional classes [10]. By implementing Logistic Regression, Random Forest, and Naive Bayes classifiers with optimized preprocessing pipelines and Term Frequency-Inverse Document Frequency vectorization, this work demonstrates that classical approaches can achieve robust performance for fine-grained emotion classification tasks. The evaluation framework encompasses multiple performance metrics including accuracy, precision, recall, F1-score, and cross-validation analysis, providing comprehensive insights into model behavior and classification patterns across different emotional categories.

2. RELATED WORK

The field of emotion detection in text has evolved significantly over the past two decades, progressing from rule-based systems and lexicon approaches to sophisticated machine learning and deep learning methodologies. Early research in affective computing established theoretical foundations for computational emotion recognition, drawing upon psychological models such as Ekman's basic emotions and Russell's circumplex model to define categorical and dimensional representations of emotional states [11]. These psychological frameworks have informed the design of emotion taxonomies and annotation schemes used in computational studies, though debates continue regarding the most appropriate emotional representation for natural language processing applications [12].

Lexicon-based approaches represented the first wave of automated emotion detection systems, relying on manually curated dictionaries that associate words and phrases with specific emotional categories or valence-arousal-dominance ratings. The Affective Norms for English Words database and the NRC Emotion Lexicon exemplify influential resources that have enabled rule-based emotion classification through keyword matching and aggregation strategies [13]. While lexicon methods offer interpretability and require no training data, they suffer from limited coverage of domain-specific vocabulary, inability to handle context-dependent emotional expressions, and challenges in processing figurative language such as sarcasm and irony [14].

Traditional machine learning approaches emerged as researchers recognized the limitations of purely lexical methods and sought data-driven solutions capable of learning emotional patterns from annotated corpora. Support Vector Machines gained prominence for text classification tasks due to their effectiveness in high-dimensional feature spaces and strong generalization capabilities [15]. Studies demonstrated that linear SVMs combined with TF-IDF features could achieve competitive performance for emotion classification while maintaining computational efficiency. Random Forests and ensemble methods have similarly shown promise by combining multiple decision trees to capture complex non-linear relationships in textual features while providing robustness against overfitting [16].

Naive Bayes classifiers, despite their simplifying independence assumptions, have demonstrated surprisingly strong performance for text classification tasks including emotion detection. The multinomial variant proves particularly suitable for document classification problems where features represent term frequencies, offering computational efficiency and interpretability that facilitate deployment in resource-constrained environments [17]. Comparative studies have shown that proper feature engineering and preprocessing can enable Naive Bayes to approach or exceed the performance of more complex algorithms for certain emotion detection tasks [8].

Feature engineering has played a crucial role in the success of traditional machine learning approaches for emotion detection. Beyond basic bag-of-words and TF-IDF representations, researchers have explored diverse feature sets including part-of-speech patterns, syntactic dependencies, sentiment lexicon scores, negation handling, and emotion-specific word embeddings [18]. N-gram features capturing multi-word expressions have proven particularly valuable for emotion classification, as emotional content often manifests through phrasal constructions rather than individual words. Character-level n-grams provide additional robustness to spelling variations and out-of-vocabulary terms common in social media text [19].

The emergence of word embeddings revolutionized feature representation in natural language processing by enabling dense vector representations that encode semantic relationships learned from large text corpora. Word2Vec and GloVe embeddings have been widely adopted for emotion detection tasks, either as standalone features or in combination with traditional representations [20]. These distributed representations capture semantic similarities between emotionally related terms, enabling models to generalize beyond exact keyword matches. However, static word embeddings fail to account for polysemy and context-dependent meaning, motivating the development of contextualized representations [21].

Deep learning approaches have achieved remarkable success in emotion detection tasks through the development of neural architectures capable of automatically learning hierarchical feature representations. Convolutional Neural Networks have been applied to text classification by treating sentences as sequential data and learning local patterns through convolution operations [22]. Recurrent Neural Networks, particularly Long Short-Term Memory and Gated Recurrent Unit variants, explicitly model sequential dependencies in text, capturing contextual information crucial for understanding emotional content [23]. Attention mechanisms further enhance these architectures by enabling models to focus on emotionally salient words and phrases while processing input sequences [24].

Transformer-based models represent the current state-of-the-art for numerous natural language processing tasks including emotion detection. BERT and its variants leverage bidirectional self-attention mechanisms and masked language modeling pretraining to learn rich contextual representations that encode both semantic and syntactic information [7]. Fine-tuning pretrained BERT models on emotion-labeled datasets has yielded exceptional performance across multiple benchmarks, often surpassing human-level agreement in emotion annotation tasks [25]. RoBERTa, ALBERT, and DistilBERT variants offer improvements in training efficiency, model compression, and computational requirements while maintaining strong performance [26].

Domain-specific challenges in emotion detection have motivated specialized research addressing particular textual characteristics and application contexts. Social media text presents unique difficulties including informal language, spelling variations, abbreviations, emoticons, and hashtags that carry emotional information [27]. Researchers have developed preprocessing strategies and domain-adapted models to handle these characteristics while preserving emotionally relevant features. Multimodal emotion recognition integrates textual analysis with acoustic and visual information from audio and video data, enabling more comprehensive emotion understanding in multimedia content [28].

Evaluation methodologies for emotion detection systems have evolved to address the inherent subjectivity in emotion annotation and the class imbalance common in real-world datasets. Inter-annotator agreement measures such as Cohen's kappa and Fleiss' kappa quantify the reliability of emotion labels, though substantial disagreement often exists even among expert annotators [29]. Researchers have explored probabilistic annotation frameworks that preserve label uncertainty and multi-label classification approaches that accommodate the simultaneous expression of multiple emotions in text. Class imbalance mitigation strategies including oversampling, undersampling, and class-weighted loss functions have been investigated to improve model performance on minority emotion categories [30].

Benchmark datasets have played a crucial role in advancing emotion detection research by providing standardized evaluation protocols and enabling systematic comparison of different approaches. The SemEval shared tasks have fostered community-wide collaboration on emotion analysis challenges, establishing common evaluation frameworks and promoting reproducible research [18]. The GoEmotions dataset represents a significant contribution by providing fine-grained emotion annotations for Reddit comments across 27 emotional categories, enabling research on subtle emotional distinctions. However, concerns regarding annotation quality, dataset bias, and limited ecological validity continue to motivate development of improved data collection and annotation methodologies.

The present work contributes to this rich research landscape by conducting a systematic comparison of classical machine learning algorithms for multi-class emotion detection across eleven emotional categories. While much recent research has focused on deep learning approaches, this study demonstrates that properly configured traditional methods can achieve competitive performance while offering advantages in computational efficiency, interpretability, and deployment simplicity. The comprehensive evaluation framework encompassing accuracy, precision, recall, F1-score, cross-validation analysis, and detailed confusion matrix examination provides insights into the strengths and limitations of different algorithmic approaches for emotion classification tasks.

3. METHODOLOGY

3.1. Dataset Description

The experimental investigation employs a comprehensive emotion detection dataset comprising approximately 106,000 sentences, each annotated with its corresponding emotional label [41]. This dataset represents a balanced corpus specifically designed for natural language processing tasks utilizing transformer-based architectures. The corpus was constructed through the strategic integration of three distinct emotion datasets to address the substantial data requirements characteristic of deep learning models, particularly those based on the transformer architecture. The dataset encompasses eleven distinct emotional categories, providing granular classification capability for nuanced emotion recognition tasks.

The dataset exhibits several notable characteristics that enhance its suitability for emotion detection research. All textual entries have undergone preprocessing to remove usernames and uniform resource locators, ensuring that the models focus exclusively on emotional content rather than user-specific or external reference information. Furthermore, the preprocessing protocol preserved hashtag content by removing only the hash symbol while retaining the associated text, recognizing that hashtags frequently convey significant emotional information [41]. The dataset contains no missing values, maintaining complete data integrity across all observations. This carefully curated corpus enables robust model training without the risk of overfitting that would typically occur with smaller datasets, while providing sufficient diversity to support generalization across various emotional expressions in natural language.

3.2. Preprocessing Pipeline

The preprocessing methodology implements a systematic approach to transform raw textual data into numerical representations suitable for machine learning algorithms. The process begins with label encoding, where the eleven categorical emotion labels are converted to numerical indices using scikit-learn's LabelEncoder class [42]. This transformation enables computational processing while preserving the categorical nature of the target variable throughout the analysis.

Text preprocessing constitutes a critical component of the feature extraction pipeline. Each sentence undergoes conversion to lowercase to ensure case-insensitive processing, followed by whitespace normalization to eliminate extraneous spacing. The preprocessing function retains alphanumeric characters along with basic punctuation marks including periods, commas, exclamation points, and question marks, as these elements frequently carry emotional significance in written communication [43]. The system implements a token limit of one hundred words per sentence to manage computational complexity while preserving the essential emotional content of each text sample.

Feature extraction employs Term Frequency-Inverse Document Frequency vectorization with carefully optimized parameters [44]. The vectorizer generates a feature space of three thousand dimensions, capturing the most informative terms across the corpus. The n-gram range extends from unigrams to trigrams, enabling the model to capture both individual word meanings and multi-word emotional expressions. Document frequency thresholds restrict features to terms appearing in at least three documents but no more than eighty-five percent of documents, effectively filtering both rare noise terms and overly common words that provide minimal discriminative information. The implementation includes English stop word removal and sublinear term frequency scaling to prevent highly frequent terms from dominating the feature space [45]. This vectorization strategy produces a sparse matrix representation where each sentence is encoded as a high-dimensional vector suitable for classification algorithms.

The dataset undergoes stratified partitioning into training and testing subsets using an eighty-twenty split ratio. The stratification procedure ensures that the class distribution remains consistent across both partitions, preventing potential bias in model evaluation that could arise from imbalanced sampling [42]. The training subset, comprising approximately 84,800 samples, supports model parameter estimation, while the testing subset of approximately 21,200 samples enables unbiased performance evaluation on previously unseen data.

3.3. Classification Models

The experimental framework incorporates three distinct machine learning algorithms, each selected for its unique approach to classification and proven effectiveness in text analysis tasks. Logistic Regression serves as a strong linear baseline, implementing multinomial classification with L2 regularization and a regularization parameter of 1.5 [46]. The model employs balanced class weights to account for any residual class imbalances and utilizes the limited-memory Broyden-Fletcher-Goldfarb-Shanno optimization algorithm with a maximum of two thousand iterations to ensure convergence. The multinomial formulation enables direct multi-class probability estimation without requiring one-versus-rest decomposition.

Random Forest classification implements an ensemble of one hundred fifty decision trees with a maximum depth of thirty levels [47]. The algorithm employs bootstrap aggregation to reduce variance while maintaining low bias through the use of deep trees. Minimum sample requirements of four samples for splitting and two samples for leaf nodes prevent excessive overfitting while allowing the model to capture complex decision boundaries. The Random Forest approach proves particularly effective for text classification due to its ability to handle high-dimensional sparse features and its inherent feature importance estimation capabilities.

Naive Bayes classification implements a probabilistic approach based on Bayes' theorem with strong independence assumptions between features [48]. The multinomial variant proves particularly suitable for text classification tasks where features represent term frequencies. The smoothing parameter alpha is set to 0.1, applying Laplace smoothing to handle zero probabilities for unseen term-class combinations. Despite its simplifying assumptions, Naive Bayes often achieves competitive performance in text classification while offering computational efficiency and interpretability through its explicit probability modeling.

All models undergo training using parallel processing capabilities where applicable to expedite the experimental workflow. The training process includes comprehensive timing measurements to assess computational efficiency alongside predictive performance. Each model's hyperparameters were selected based on preliminary experiments and established best practices for text classification tasks, balancing model complexity with generalization capability.

3.4. Evaluation Metrics

The evaluation framework implements a comprehensive suite of metrics to assess model performance across multiple dimensions, ensuring robust characterization of classification capability. Accuracy serves as the primary metric, quantifying the proportion of correct predictions across all emotion categories. While accuracy provides an intuitive overall performance measure, the evaluation extends beyond this single metric to capture nuanced aspects of model behavior.

Precision, recall, and F1-score are calculated for each emotion category and aggregated using weighted averaging to account for class frequency in the test set [42]. Precision measures the proportion of correct predictions among all instances predicted as a particular emotion, quantifying the model's ability to avoid false positive errors. Recall measures the proportion of actual instances of each emotion that the model successfully identifies, quantifying completeness of detection. The F1-score represents the harmonic mean of precision and recall, providing a balanced measure that penalizes extreme trade-offs between these complementary metrics. The weighted averaging scheme ensures that the aggregate metrics appropriately reflect performance across all emotion categories proportional to their representation in the test data.

Cross-validation analysis employs a five-fold stratified approach to assess model stability and generalization capability [49]. The training data is partitioned into five equal subsets while maintaining class distribution within each fold. Each model undergoes training on four folds and evaluation on the remaining fold, with this process repeated five times such that each fold serves once as the validation set. The cross-validation procedure yields five performance estimates for each model, from which mean and standard deviation statistics are computed. The mean cross-validation score provides an estimate of expected performance on unseen data, while the standard deviation quantifies prediction stability across different training subsets. Models exhibiting low standard deviation demonstrate consistent behavior regardless of specific training samples, suggesting robust generalization capability.

The confusion matrix provides detailed insight into classification patterns by tabulating predicted labels against true labels for all test samples [50]. This matrix reveals not only overall accuracy but also specific misclassification patterns, identifying which emotion pairs are most frequently confused. The analysis includes both absolute confusion matrices displaying raw prediction counts and normalized confusion matrices presenting proportions, facilitating interpretation across emotion categories with varying frequencies. Systematic patterns in the confusion matrix often reveal semantically meaningful relationships, such as higher confusion rates between emotions of similar valence or intensity.

Per-class performance metrics enable identification of emotions that present particular classification challenges. Computing precision, recall, and F1-score separately for each emotion category reveals whether the model performs uniformly across all emotions or exhibits strengths and weaknesses for specific categories. This granular analysis proves valuable for understanding model limitations and guiding future improvements, such as targeted data collection for underperforming categories or specialized feature engineering for emotionally ambiguous expressions.

Training time measurements quantify the computational efficiency of each algorithm, providing practical insights for deployment scenarios where training time constraints may influence model selection [42]. While predictive performance remains the primary consideration, computational requirements often play a crucial role in determining feasibility for applications requiring frequent retraining or operation under resource constraints. The comprehensive evaluation framework thus balances multiple considerations, enabling informed model selection based on the specific requirements and constraints of the target application.

4. Results and Discussion

The experimental evaluation of three machine learning algorithms for emotion detection yielded distinct performance characteristics across multiple evaluation metrics, as illustrated in Figure 1. Logistic Regression emerged as the superior model, achieving an overall accuracy of 79.90% on the test dataset comprising approximately 21,200 samples. This performance substantially exceeded both Random Forest, which attained 75.32% accuracy, and Naive Bayes, which achieved 69.01% accuracy. The consistent superiority of Logistic Regression across all evaluation metrics demonstrates its effectiveness for multi-class emotion classification in text data.

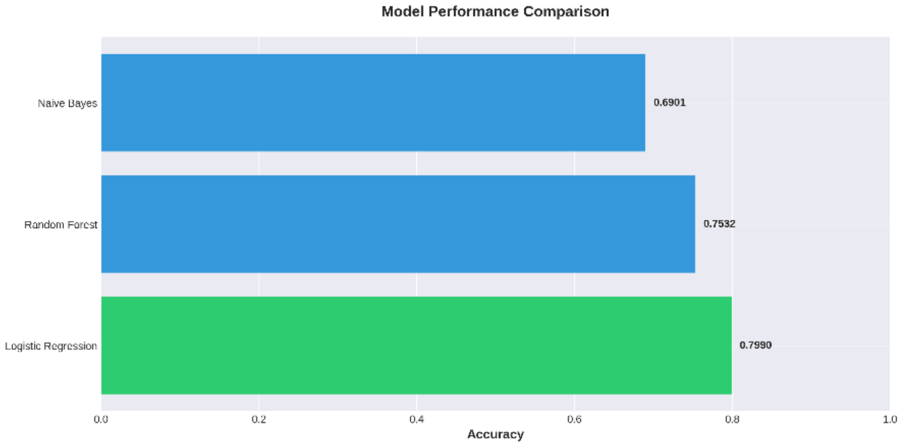

Figure 1. Comparative analysis of model accuracy across three classification algorithms. Logistic Regression achieved the highest performance at 79.90%, followed by Random Forest and Naive Bayes.

The comprehensive performance analysis presented in Figure 2 reveals that Logistic Regression not only achieved the highest accuracy but also demonstrated balanced performance across precision, recall, and F1-score metrics. The model attained a weighted precision of 81.18%, indicating strong reliability in its positive predictions, while maintaining a recall of 79.90%, demonstrating effective identification of emotion instances across all categories. The F1-score of 80.27% reflects the harmonious balance between precision and recall, suggesting that the model avoids extreme trade-offs between false positives and false negatives. Random Forest exhibited notably higher precision at 85.38% compared to its recall of 75.32%, indicating a conservative prediction strategy that prioritizes accuracy over completeness. This precision-recall gap of approximately ten percentage points suggests the Random Forest model requires higher confidence thresholds before classifying instances, resulting in missed detections for ambiguous cases. Naive Bayes demonstrated the most balanced precision-recall relationship at approximately 69%, though at lower absolute performance levels than the other models.

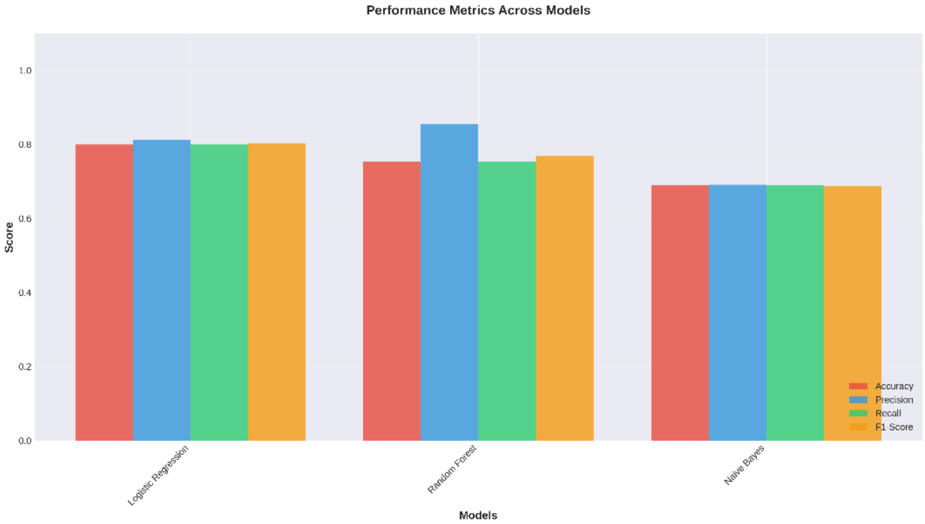

Figure 2. Performance metrics comparison shows accuracy, precision, recall, and F1-score across models. Logistic Regression demonstrates balanced performance across all metrics while Random Forest exhibits higher precision than recall.

The cross-validation analysis presented in Figure 3 provides critical insights into model stability and generalization capability. Logistic Regression achieved a mean cross-validation score of 79.4% with a standard deviation of only 0.4%, demonstrating remarkable consistency across different data partitions. This minimal variance indicates robust performance independent of specific training samples, suggesting strong generalization to unseen data. Random Forest obtained a cross-validation score of 75.2% with comparable stability at 0.4% standard deviation, while Naive Bayes achieved 67.7% with slightly higher variance at 0.5%. The tight error bars across all models indicate that performance differences reflect genuine algorithmic characteristics rather than random variation or overfitting to particular data subsets.

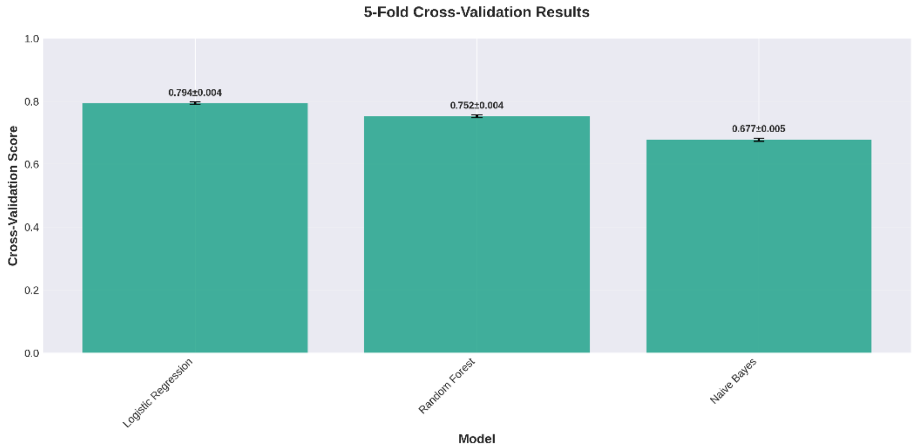

Figure 3. Five-fold cross-validation results with error bars indicating standard deviation across folds. All models demonstrate stable performance with minimal variance, confirming robust generalization capability.

The detailed confusion matrix analysis for Logistic Regression, depicted in Figures 4 and 5, reveals nuanced patterns in classification performance across the eleven emotion categories. The normalized confusion matrix in Figure 5 demonstrates that certain emotions achieve exceptionally high classification accuracy, with enthusiasm reaching 90.45% correct classification, love at 89.15%, neutral at 83.95%, fun at 82.45%, and anger at 82.80%. These high-performing categories typically possess distinctive linguistic markers that facilitate reliable detection. Conversely, emotions such as empty, neutral, and sadness exhibit more substantial confusion with related categories. The empty emotion shows notable misclassification with neutral instances at 14.79%, reflecting the semantic similarity between these low-arousal states. Happiness demonstrates confusion with surprise at 7.60% and sadness at 7.30%, suggesting challenges in distinguishing emotions with overlapping expressive patterns.

The per-class performance analysis illustrated in Figure 6 reveals significant variation in detection capability across emotion categories. Enthusiasm emerges as the most reliably detected emotion, achieving precision, recall, and F1-scores all exceeding 90%. This exceptional performance likely stems from distinctive vocabulary associated with enthusiastic expression. Conversely, empty, neutral, and sadness present greater classification challenges, with F1-scores ranging from 65% to 70%. The lower performance for these categories may reflect their linguistic subtlety and overlap with multiple other emotional states. Anger, love, and relief demonstrate strong balanced performance with F1-scores approaching 90%, indicating the presence of characteristic linguistic features that enable reliable identification.

The computational efficiency analysis presented in Figure 7 reveals substantial differences in training requirements across algorithms. Naive Bayes demonstrates remarkable efficiency with training completion in merely 0.03 seconds, making it highly suitable for applications requiring frequent model updates or operating under severe computational constraints. Random Forest requires moderate computational resources at 4.92 seconds, representing a reasonable balance between performance and efficiency. Logistic Regression, despite achieving the highest predictive performance, demands the longest training time at 16.69 seconds due to its iterative optimization process over the high-dimensional feature space. However, this training duration remains acceptable for most practical applications, particularly given that training occurs offline and prediction latency remains minimal across all models.

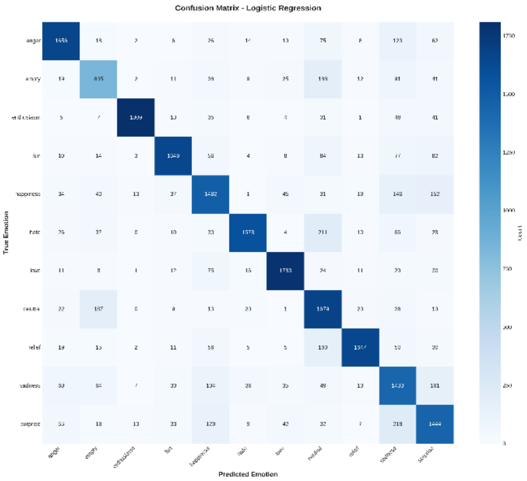

Figure 4. Confusion matrix for Logistic Regression showing absolute prediction counts across eleven emotion categories. Diagonal elements represent correct classifications while off-diagonal elements indicate misclassification patterns.

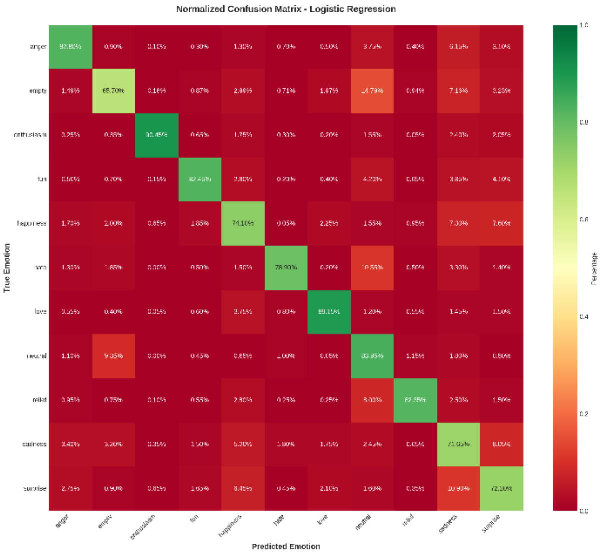

Figure 5. Normalized confusion matrix displaying classification accuracy as percentages for each emotion category. High-performing emotions include enthusiasm (90.45%), love (89.15%), and neutral (83.95%).

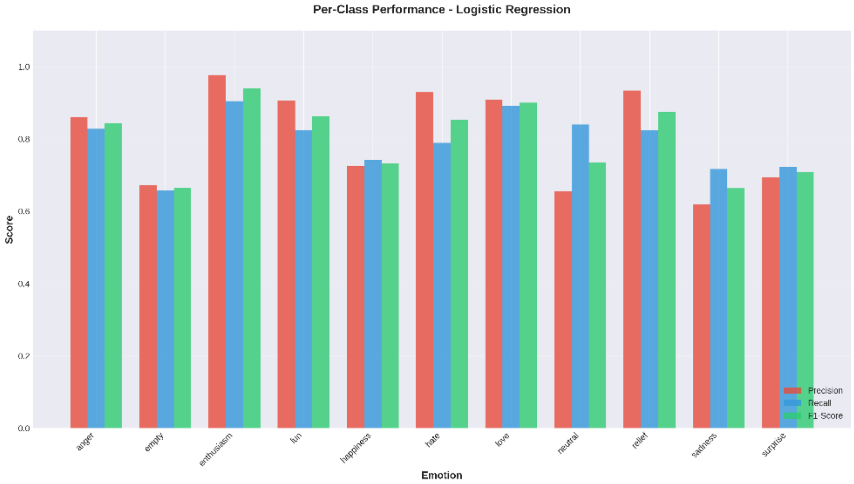

Figure 6. Per-class performance metrics showing precision, recall, and F1-score for each emotion category. Enthusiasm achieves the highest performance while empty and neutral present greater classification challenges.

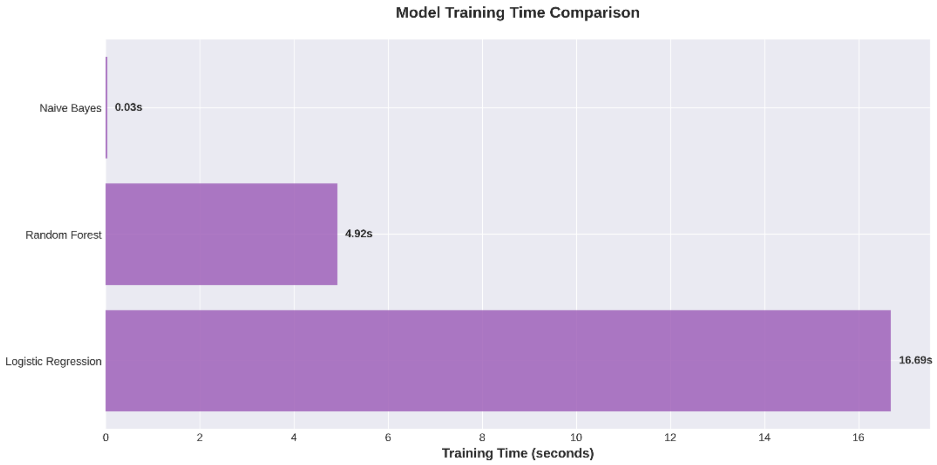

Figure 7. Training time comparison across three classification algorithms measured in seconds. Naive Bayes exhibits exceptional computational efficiency while Logistic Regression requires longer training duration.

The overall results demonstrate that classical machine learning approaches, when properly configured with optimized preprocessing and feature extraction pipelines, achieve strong performance for emotion detection tasks. Logistic Regression's success validates the hypothesis that linear models can effectively capture emotion-relevant patterns in TF-IDF transformed text features. The model's ability to achieve approximately 80% accuracy across eleven distinct emotion categories represents substantial improvement over random baseline performance of approximately 9%, demonstrating genuine learning of emotional semantics from linguistic features.

5. CONCLUSION

This study demonstrates that classical machine learning algorithms can achieve strong performance in multi-class emotion detection across eleven emotional categories using a balanced dataset of 106,000 textual instances. Among the evaluated models, Logistic Regression exhibited the highest performance, attaining an accuracy of 79.90%, and consistently outperforming both Random Forest and Naïve Bayes classifiers. Cross-validation results further confirmed the robustness and stability of the models, with Logistic Regression achieving a mean score of 79.4% and a standard deviation of only 0.4%, indicating reliable generalization across different training subsets.

Emotion-wise performance analysis revealed that categories with clearer linguistic markers such as enthusiasm, love, fun, neutral, and anger were classified with higher reliability, whereas low-arousal emotions, including sadness and empty, posed greater classification challenges due to semantic overlap. Although Logistic Regression required a longer training time compared to the other models, its superior predictive performance justifies the additional computational cost.

Overall, the findings confirm that optimized classical machine learning particularly Logistic Regression offers an accurate, interpretable, and computationally efficient solution for fine-grained emotion classification, with significant potential for practical applications in areas such as mental health monitoring, customer service analytics, social media analysis, and educational technologies.

REFERENCES

[1] S. M. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words," in Proc. 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 174-184.

[2] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," in A Practical Guide to Sentiment Analysis. Cham, Switzerland: Springer, 2017, pp. 1-10.

[3] S. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," Engineering Reports, vol. 2, no. 7, e12189, 2020.

[4] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning based text classification: A comprehensive review," ACM Computing Surveys, vol. 54, no. 3, pp. 1-40, 2021.

[5] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, 2nd ed. Cambridge, UK: Cambridge University Press, 2020.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2019, pp. 4171-4186.

[8] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in Proc. AAAI-98 Workshop on Learning for Text Categorization, Madison, WI, USA, 1998, pp. 41-48.

[9] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[10] P. Nayak, "Text Emotion Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/prajwalnayakat/text-emotion/data. [Accessed: Dec. 8, 2024].

[11] P. Ekman, "An argument for basic emotions," Cognition and Emotion, vol. 6, no. 3-4, pp. 169-200, 1992.

[12] J. A. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161-1178, 1980.

[13] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," Computational Intelligence, vol. 29, no. 3, pp. 436-465, 2013.

[14] A. Giachanou and F. Crestani, "Like it or not: A survey of Twitter sentiment analysis methods," ACM Computing Surveys, vol. 49, no. 2, pp. 1-41, 2016.

[15] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Proc. European Conference on Machine Learning, Chemnitz, Germany, 1998, pp. 137-142.

[16] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[17] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with Naive Bayes: Which Naive Bayes?" in Proc. Third Conference on Email and Anti-Spam, Mountain View, CA, USA, 2006, pp. 27-28.

[18] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 Task 4: Sentiment analysis in Twitter," in Proc. 10th International Workshop on Semantic Evaluation, San Diego, CA, USA, 2016, pp. 1-18.

[19] A. Hassan and A. Mahmood, "Deep learning approach for sentiment analysis of short texts," in Proc. 2017 3rd International Conference on Control, Automation and Robotics, Nagoya, Japan, 2017, pp. 705-710.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[21] M. E. Peters et al., "Deep contextualized word representations," in Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 2018, pp. 2227-2237.

[22] Y. Kim, "Convolutional neural networks for sentence classification," in Proc. 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 2014, pp. 1746-1751.

[23] R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in Proc. 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 2013, pp. 1631-1642.

[24] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 2016, pp. 1480-1489.

[25] N. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Transformer models for text-based emotion detection: A review of BERT-based approaches," Artificial Intelligence Review, vol. 54, no. 8, pp. 5789-5829, 2021.

[26] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[27] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in Proc. Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 2020, pp. 1644-1650.

[28] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98-125, 2017.

[29] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37-46, 1960.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[31] P. Nayak, "Text Emotion Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/prajwalnayakat/text-emotion/data. [Accessed: Dec. 8, 2024].

[32] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[33] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2019, pp. 4171-4186.

[34] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning based text classification: A comprehensive review," ACM Computing Surveys, vol. 54, no. 3, pp. 1-40, 2021.

[35] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in Proc. First Instructional Conference on Machine Learning, Piscataway, NJ, USA, 2003, pp. 133-142.

[36] C. Zhang and Y. Ma, Ensemble Machine Learning: Methods and Applications. New York, NY, USA: Springer, 2012.

[37] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[38] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in Proc. AAAI-98 Workshop on Learning for Text Categorization, Madison, WI, USA, 1998, pp. 41-48.

[39] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," Statistics Surveys, vol. 4, pp. 40-79, 2010.

[40] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," arXiv preprint arXiv:2008.05756, 2020.