Emotion Prediction Using Natural Language Processing: A Performance Evaluation of Supervised Machine Learning Models for Classification Tasks

Dennis S. Nava, Roman B. Villones

La Consolacion University Philippines

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150100052

Received: 22 January 2026; Accepted: 27 January 2026; Published: 06 February 2026

ABSTRACT

This study discusses the usage of machine learning algorithms in the detection of emotions. It helps to fill the gap between the human understanding of emotions and the artificial intelligence being developed, both in relation to technological progress and human-machine interactions. Applying the KDD process to NLP systematically involves identifying, preprocessing, mining patterns, and interpreting textual data. Techniques of Data Mining, like Logistic Regression, Linear SVC, and Multinomial Naïve Bayes, have been applied for the purpose, accompanied by metrics like Classification Reports for Precision, Recall, F1-score, and Accuracy and k-fold cross-validation for a very robust and accurate analysis of the unstructured text domain. Logistic Regression remained the most scoring model of 90%. Its score now was a few percentage points smaller than its training score. As expected, this time Linear SVC found itself in second place with an 88%, and Multinomial Naïve Bayes also stayed in third place on 86%. With cross-validation applied, Logistic Regression proves to be the most reliable model for this NLP study. It scores high and is quite generalizable and stable, thus suitable for deployment in scenarios where robustness and consistency are essential. Potential further improvements to such models could involve hyperparameter and fine-tuning. By trying them out on much larger and varied datasets, they can be put through several checks for testing capabilities and limits within applications. One other extension involves ensemble techniques to build even stronger, more reliable solutions by merging different model strengths together.

Keywords – Logistic Regression, Linear SVC, and Multinomial Naïve Bayes, Machine Learning, Natural Language Processing.

INTRODUCTION

The ability to predict emotions is becoming a highly essential focus of research in the 21st century, cut across psychology, computer science, and artificial intelligence disciplines. Emotions take a central place in human decision-making, communication, and even well-being. Understanding and predicting emotions is also relevant to applications in areas like mental health, human-computer interaction, marketing, and education. It thus emerges as one of the key areas of contemporary science and technology. In the study of Kalateh et al. (2024) Probably, the biggest obstacle in emotion study is the hundreds of definitions, which are prone to terminological confusion. The paper was based on the compilation of 92 definitions and 9 skeptical statements into a single paper, organized into 11 categories based on different emphases on emotional phenomena or theoretical issues. In addition of Mahfoudi et al. (2022) When people identify emotion as a key domain for computing, they are emphasizing its possible contribution to enhancing human-computer interactions such as emotion and sentiment analysis, or affective computing.

According to Al Maruf et al. (2024) Machine learning techniques are a superior choice for emotion detection tasks as they enable the model to classify emotions even when texts indirectly reference emotions. Good results, however, require good lexical resources being used as features in machine learning algorithms.

Emotions can be portrayed through many observable modes of communication, for example facial expression, gestures, words and texts. The perception of emotions within a text document is essentially a content-based classification problem that brings together concepts from both Natural Language Processing and Machine Learning.

This research aims to study different machine learning models for emotion detection implementations and challenges. It seeks to contribute to the ever-growing body of knowledge that connects the gap between humans emotional understanding and artificial intelligence through examining current methodologies and their implications. Such research not only improves technological capabilities but also encourages a better insight into human emotional states.

LITERATURE REVIEW

The study of Kišjuhas (2024) recent research on emotions has clearly distinguished into four discrete areas: the influence of culture, biological components, internal feelings, and outward expressions of emotional activity. Along with this bifurcation, there is also an increasing ascendency of those researchers who investigate only the cultural and external elements of emotions. This trend may fail to recognize that these elements collaborate in a dynamic relationship that would be necessary to achieve a fully integrative model of emotional processes.

According to Cowen and Keltner (2021) A critique of this study is its point that a proper classification for today's affective science cannot fit all the terms placed on the same rank as traditional concepts such as emotion, anger, or fear anymore; therefore, their current frames cannot be very representative of real-life human experiences under complex conditions found today in research on emotions. As explained by Guo et al. (2024) Machine learning algorithms are strongly dependent on the data processing and the volume of the datasets used. In some situations, a machine learning model fails to pick up implicit features or subtle features of text. This might create the need for better pre-processing techniques as well as improvement on feature extraction methods to enhance the accuracy of these models when handling complex data texts.

According to Nandwani and Verma (2021) Emotion detection on text has become increasingly popular in recent years because of its wide-ranging implementations in sectors such as marketing, political science, psychology, human-computer interaction, and artificial intelligence. The availability of extensive textual data, particularly opinionated and self-expressive content, has accelerated interest in this domain. This growth emphasizes the field's significance and capability to reshape how emotional insights are extracted and used across multiple sectors. In addition of Hung and Alias (2023) Emotions are basic part of human existence and comprise a significant share of decision-making processes. While computers have long been used to support decision-making, they have traditionally relied on objective, factual data. Lately, there has been an increasing interest among researchers in the detection of subjective information, especially in blogs and online social media. This is a shift towards recognition of emotions and sentiments content as critical factors in analyzing and advancing human-computer interactions.

Emotion detection is a subset of sentiment analysis, a method that identifies and analyzes emotions. With the rise of Web 2.0, text mining and analysis have come to the forefront of institutional success, enabling service providers to offer customized services for their customers. The availability of data and the significant advantages it offers have driven extensive research in text mining and analysis, reflecting its increasing significance in refining customer experience and decision-making processes as explained by Wankhade et al. (2022).

METHODOLOGY

Knowledge Discovery in Databases (KDD Process)

The KDD Process is another term for Knowledge Discovery in Databases, which is the structured methodology for extracting useful, valid, and actionable knowledge gained from extensive datasets It is comprised of several stages, namely data selection, preprocessing, transformation, data mining, and interpretation/evaluation of the discovered knowledge according to Aldoseri et al. (2024).

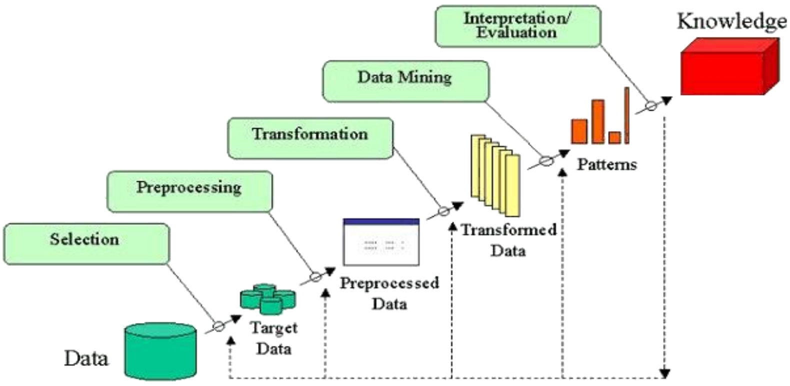

Figure 1. Knowledge Discovery in Databases

Figure 1 shows an application of the KDD process on Natural Language Processing (NLP) focuses on identification of textual data, preprocessing, pattern mining, and interpretation of results with better insights to be gained. This systematic approach assures one systematically investigates the unstructured rich domain of text for robust outcomes.

Selection

The datasets have 422,746 rows indexed sequentially from 0 to 422,745 and it has 2 columns named "sentence" and "emotion". Both the columns have non-null values which means that there is no missing data in this dataset. This indicates that the column "sentence" is holding text data, and the column "emotion" holds the corresponding labels of emotions ('fear', 'sad', 'love', 'joy', 'suprise', 'anger') that describe the emotional situation of each sentence.

Preprocessing

First, after cleaning the dataset, 6,623 duplicate entries were found and removed to make the data valid and reliable.

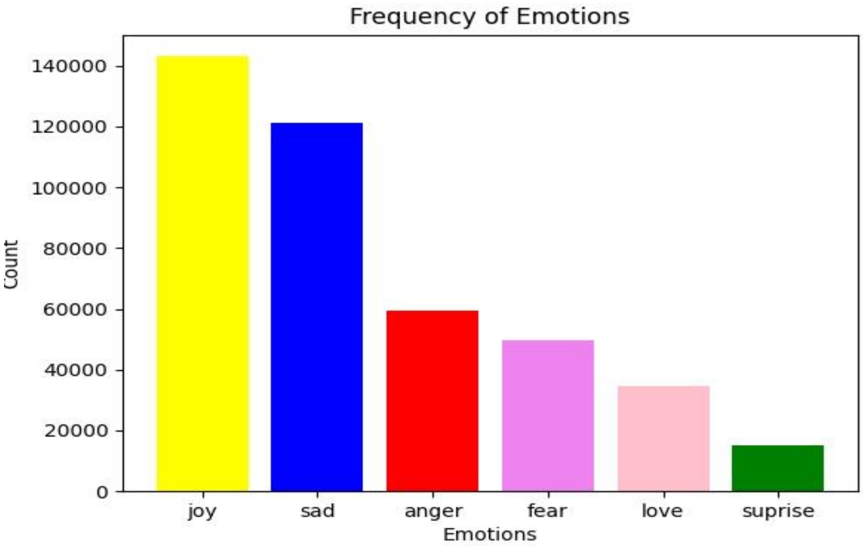

Figure 2. Frequency of Emotions

Figure 2 shows t the dataset counts of different emotions, which are distributed as follows: 'joy' is the most frequent emotion with 140,779 occurrences, followed by 'sadness' with 120,989 occurrences. 'Anger' occurs 57,235 times, 'fear' 47,664 times, 'love' 34,497 times, and 'surprise' is the least frequent with 14,959 occurrences.

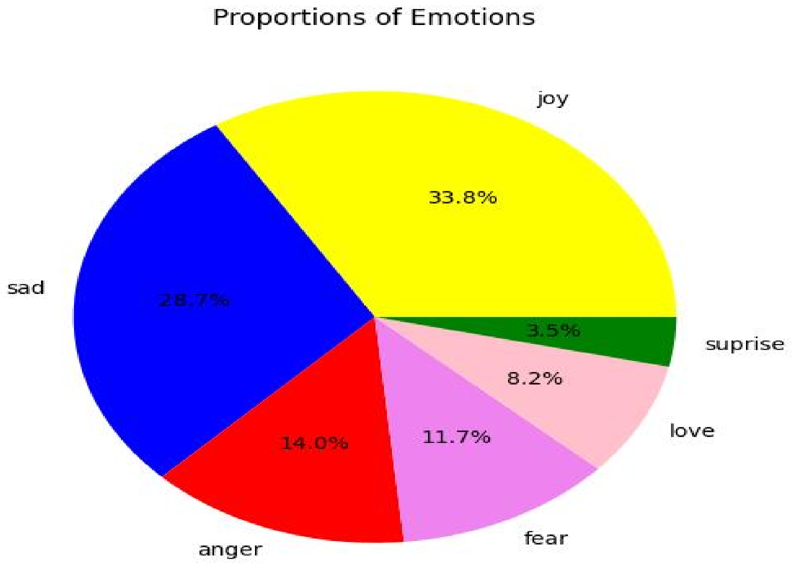

Figure 3. Proportions of Emotions

Figure 3 represents the proportional distribution of different emotions in the dataset. The largest chunk is 'joy', which amounts to 33.8% of the overall data. The second most dominant emotion is 'sadness', which accounts for 29.1%. Other emotions are represented as follows: 'anger' at 13.8%, 'fear' at 11.5%, 'love' at 8.3%, and 'surprise' at 3.6%, which is the least frequent emotion. The chart gives a visual sense of how 'joy' and 'sadness' dominate the rest of the emotions.

Transformation

In the transformation of these datasets is achieved using the `CountVectorizer` from `sklearn` libraries, which transforms text data in the "sentence" column into a form that would be usable in machine learning models. Here, it converts sentences into a sparse matrix of token counts. Each row here represents a sentence, and each column corresponds to a unique word or token found in the dataset. The values in the matrix represent how often each word is repeated in its corresponding sentence. This technique is called the bag-of-words representation, which is vital in text preprocessing: it makes the textual data representable numerically while maintaining word occurrence information.

The training data applies the `fit_transform` function to learn the vocabulary from the "sentence" column and transform the text into a sparse matrix of token counts simultaneously. This ensures that the model is aware of the structure and frequency of words in the training dataset.

For the test data, the function is `transform`, however. The latter transforms test sentences into a form of sparse matrix based on learned vocabulary, just like with the training data and does not alter or add tokens. That will ensure the representational consistency for the train data and the test data that makes the difference when evaluating an actual model.

Data Mining

Logistic Regression

According to Almohimeed et al. (2023) Logistic regression changes the emphasis from outcome value modeling to a consideration of relative probabilities (odds) for certain categories of results. It turns out that, over most of its range, the natural logarithm of such odds is linear, allowing techniques developed for linear models to be applied to logistic regression. The approach focuses on how logistic regression can be viewed as a method for modeling categorical outcomes using linear techniques adapted for probabilistic interpretation.

Linear Support Vector Classifier

As explained by Dhanya et al. (2022) Linear SVM analysis seeks to find a one-dimensional hyperplane, or a line, that separates data points according to their target categories. There are infinitely many possible lines, but the problem is finding the best one. This is a critical decision because the best line maximizes the margin between categories, thus providing better generalization and classification performance.

Multinomial Naïve Bayes

As mentioned by Álvarez-Carmona et al. (2022) The effectiveness of the Naïve Bayes algorithm in sentiment classification has been proven to be quite high, even surpassing other algorithms for performance. Multinomial Naïve Bayes is chosen when speedy and simple text classification processing is required. The work is based on the idea of multinomial distribution, and hence, it was identified to work nicely with the text. Previously, the main usages of the bag-of-words feature extraction approach were known for compatibility with the approach and good capturing ability of textual patterns.

Classification Report (Precision, Recall, F1-score, and Accuracy)

Precision calculates the percentage of true positive predictions out of the total number of instances that have been predicted to be positive. Recall calculates the percentage of actual positive instances that are correctly classified. The harmonic means of precision and recall, known as the F1-score, is often a weighted average of both. Accuracy refers to the percentage of correct predictions with respect to the total instances evaluated. Collectively, these metrics give an all-rounded appraisal of the performance of a model regarding correctness and completeness based on Khan et al. (2024).

k-Fold Cross Validation

According to Tapeh and Naser (2023) KCV or K-fold cross-validation is applied in model selection, and error estimation for classification task. The method splits a dataset into k subsets and uses them alternately as training and validation sets to evaluate the ability of the model to generalize. This is an iterative process, ensuring that the model is trained and validated on all data available, reducing the risk of overfitting and increasing stability.

RESULTS

Interpretation/ Evaluation

The evaluation metrics for the Logistic Regression, Linear SVC, and Multinomial Naïve Bayes model depict its performance across six emotion classes. In most of the classes, the developed models were able to achieve a good level of precision, recall, f1-score, and accuracy.

Logistic Regression

Emotion	Precision	Recall	F1-Score	Support
Anger	0.92	0.93	0.92	57,235
Fear	0.87	0.89	0.88	47,664
Joy	0.94	0.94	0.94	140,779
Love	0.83	0.82	0.82	34,497

Sad	0.96	0.95	0.95	120,989
Surprise	0.8	0.79	0.79	14,959
Accuracy			0.92	416,123
Macro Avg	0.89	0.88	0.89	416,123
Weighted Avg	0.92	0.92	0.92	416,123

Table 1. Logistic Regression Classification Report

Table 1 shows the Classification Report of Logistic Regression Model and It results for 'anger' a 0.92. Similarly, 'fear' had slightly less metrics, i.e., with 0.88. The model did very great work for 'joy' as well as 'sadness' with both being 0.94 and 0.95 respectively. For 'love', the numbers were slightly below that with 0.82. Here 'surprise' had a low performance at 0.79. Overall, the model performed on average at the level of about 0.92 or even 92 %.

Linear Support Vector Classifier

Emotion	Precision	Recall	F1-Score	Support
Anger	0.94	0.94	0.94	57,235
Fear	0.89	0.90	0.89	47,664
Joy	0.95	0.95	0.95	140,779
Love	0.84	0.84	0.84	34,497
Sad	0.96	0.96	0.96	120,989
Surprise	0.82	0.81	0.81	14,959
Accuracy			0.93	416,123
Macro Avg	0.90	0.90	0.90	416,123
Weighted Avg	0.93	0.93	0.93	416,123

Table 2. Linear SVC Classification Report

Table 2 shows the Classification Report of Linear SVC Model, for instance, 'anger' achieves a score of 0.94, showing excellent accuracy for detecting this emotion. Then, 'fear' had relatively slightly lower metrics of 0.89. In the case of 'joy' and 'sadness', the model performed very well again with scores of (0.95 and 0.96). However, in 'love', the scores are a bit low with 0.84, whereas 'surprise' has a slight score of 0.81, implying some problem with its detection. The overall accuracy of the model is 0.93, which means it correctly classifies 93% of all instances, respectively it is reflecting strong overall performance.

Multinomial Naïve Bayes

Emotion	Precision	Recall	F1-Score	Support
Anger	0.92	0.88	0.90	57,235
Fear	0.86	0.85	0.86	47,664
Joy	0.88	0.94	0.91	140,779
Love	0.85	0.7	0.76	34,497
Sad	0.9	0.95	0.92	120,989
Surprise	0.89	0.48	0.63	14,959
Accuracy			0.89	416,123
Macro Avg	0.88	0.80	0.88	416,123
Weighted Avg	0.89	0.89	0.89	416,123

Table 3. Multinomial Naïve Bayes Classification Report

Table 3 shows the Classification Report of Multinomial Naïve Bayes, the model performs well for certain emotions like 'sad' with 0.92, 'joy' with 0.91, and 'anger' with 0.90 indicating high accuracy in identifying these emotions. And the `fear` emotion scores of 0.86 evidently average for the results. However, the score for 'love' is comparatively lower at 0.76 and 'surprise' scores 0.63, indicating challenges in class prediction. The accuracy of the model is 0.89 overall, meaning it can correctly classify 89% of all instances in the dataset.

Overall Performances

Rank	Model	Scores
1	Linear Support Vector Classifier	93%
2	Logistic Regression	92%
3	Multinomial Naïve Bayes	89%

Table 4. Ranking of Performances of the Model

Table 4 compares the ranking of performance of three machine learning models based on their evaluation scores.

After the k-Fold Cross Validation

The data splitting is used in a 10-fold cross-validation procedure. It is a strong procedure for evaluating the performance of a machine learning model. It takes a data set and splits it into 10 subsets, commonly called "folds." On every iteration of cross-validation, the algorithm takes one-fold as its test set, leaving nine to use together in one combined fold. Therefore, it gives equal representation for the training and the test data to the point of over testing, it exhausts one test instance on average but exercises all of nine on training.

Logistic Regression

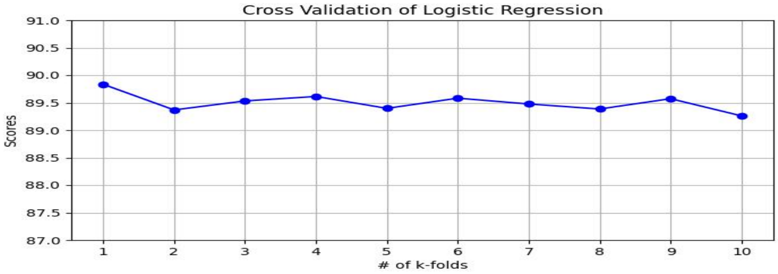

Figure 4. 10-fold Cross Validation Result for Logistic Regression

Figure 4 shows the Logistic Regression score before cross-validation is 0.92; this would reflect the accuracy when the model has been trained and tested on the same single split of a dataset, likely not cross-validation; after 10-fold cross-validation the overall average drops slightly to 0.89502382 or round up to 0.90, that is much more realistic and trustworthy because it allows for the account of variability within the data.

Linear Support Vector Classifier

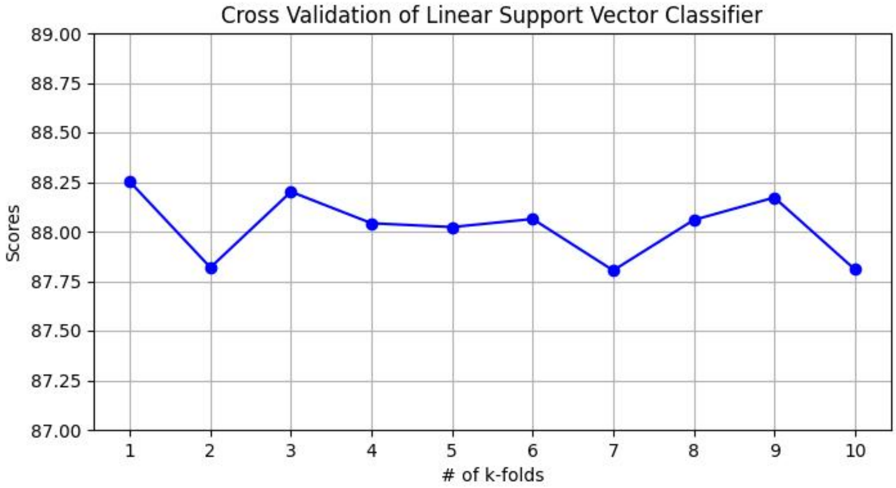

Figure 5. 10-fold Cross Validation Result for Linear SVC

Figure 5 shows the Linear SVC score before cross-validation is 0.93, which may refer to the accuracy in a single train-test split evaluation of the model, probably not cross-validated. Whereas after doing 10-fold cross-validation, the average accuracy falls to 0.880241966 or rounded to 0.88, which is closer to the estimation of the performance of the model.

Multinomial Naïve Bayes

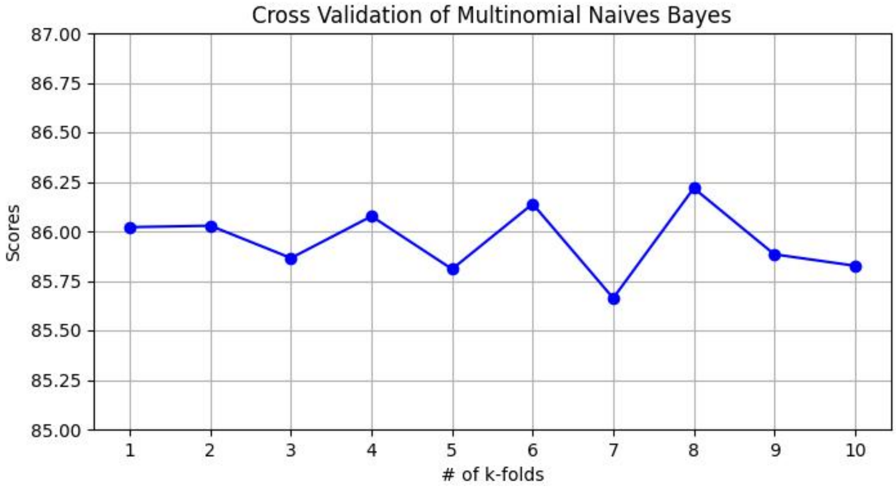

Figure 6. 10-fold Cross Validation Result for Linear SVC

Figure 6 shows the Multinomial Naïve Bayes score before cross-validation is 0.89, which gives the accuracy attained in a single train-test split likely without cross-validation. On the other hand, after performing 10-fold cross-validation, the average accuracy falls to 0.859529514 or roundup by 0.86 that gives a better estimate of the model's generalization ability on the unseen data.

Overall k-Fold Cross Validation Results

Rank	Model	Before	After
1	Logistic Regression	92%	90%
2	Linear Support Vector Classifier	93%	88%
3	Multinomial Naïve Bayes	89%	86%

Table 5. Ranking by k-fold Cross Validation results

Table 5 shows the ranking of three machine learning models before and after performing 10-fold cross-validation.

DISCUSSION

Among the machine learning models, Linear SVC was the top scorer with a score of 93%, followed by Logistic Regression at 92%. It suggests that this is a strong performer, though just slightly less effective than the model that topped the list. The third best scorer was Multinomial Naïve Bayes at 89% but still provided a respectable performance. The minimal difference in scores indicates that all three models work well, although the Linear SVC seems to be the best fit for this problem.

At the beginning, the Linear SVC had a score of 93%, then Logistic Regression 92%, and Multinomial Naïve Bayes at 89%. The ranking changed after applying 10-fold cross-validation. Logistic Regression was the leading model at 90%, although this is lower than its score at the start. The Linear SVC performed more significantly in comparison, scoring it down to 88% and taking it to second place. Multinomial Naïve Bayes was able to hold on to third place with a final score of 86%.

The shift in performance after cross-validation shows the significance of this technique in providing a more detailed evaluation of model effectiveness as it minimizes overfitting and better reflects the model true predictive capabilities. Logistic Regression is now the most reliable model suited for this task based on the validated results.

CONCLUSIONS AND RECOMMENDATIONS

This study compared three machine learning algorithms namely, Logistic Regression, Linear SVC, and Multinomial Naïve Bayes and using 10-fold cross-validation. Initially, the Linear SVC gained the highest score, followed by Logistic Regression, and then by Multinomial Naïve Bayes. However, after 10-fold cross-validation, the rankings of the models have changed. Showing that Logistic Regression is the best model with an end score of 90%. The Linear SVC score decreased to 88%, and Multinomial Naïve Bayes remained at the third position with 86%. These results indicate the significance of cross-validation in giving a stronger and precise performance evaluation for identifying the most reliable algorithm for real-world applications.

Based on the validated results, the Logistic Regression is recommended as the most reliable model for this Natural Language Processing study. The stable performance after cross-validation indicates strong generalizability, making it suitable for implementation in scenarios where robustness is crucial.

Future research should consider different techniques including hyperparameter and fine-tuning for additional performance improvements from these models. It may also give a broader and more comprehensive perspective of what their capabilities are. What the limitations are in these models for certain applications and situations if applied more substantial and varied datasets. And lastly, the approach to ensemble techniques to pool strength of various models that would give much higher accuracy than before.

Ethical Approval

This study did not require ethical approval because it used secondary data derived from the academic evaluation records of students from a private school. The dataset was accessed in anonymized and aggregated form, with no personal, sensitive, or identifiable information included. As a result, the research does not involve direct human participation and poses no ethical risk, in accordance with standard research ethics guidelines.

Conflict of Interest

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

Statement

The datasets used in this study are not publicly available due to institutional and confidentiality restrictions of the private school from which the data were obtained. However, the data may be made available upon reasonable request and subject to approval by the corresponding private institution.

REFERENCES

1. Al Maruf, A., Khanam, F., Haque, M. M., Jiyad, Z. M., Mridha, M. F., & Aung, Z. (2024). Challenges and opportunities of text-based emotion detection: A survey. IEEE access, 12, 18416-18450.
2. Aldoseri, A., Al-Khalifa, K. N., & Hamouda, A. M. (2024). AI-powered innovation in digital transformation: Key pillars and industry impact. Sustainability, 16(5), 1790.
3. Álvarez-Carmona, M. Á., Aranda, R., Rodríguez-Gonzalez, A. Y., Fajardo-Delgado, D., Sánchez, M. G., Pérez-Espinosa, H., Martínez-Miranda, J., Guerrero-Rodríguez, R., Bustio-Martínez, L., & Díaz-Pacheco, Á. (2022). Natural language processing applied to tourism research: A systematic review and future research directions. Journal of king Saud university-computer and information sciences, 34(10), 10125-10144.
4. Almohimeed, A., Saad, R. M., Mostafa, S., El-Rashidy, N. M., Farrag, S., Gaballah, A., Elaziz, M. A., El-Sappagh, S. & Saleh, H. (2023). Explainable artificial intelligence of multi-level stacking ensemble for detection of Alzheimer's disease based on particle swarm optimization and the sub-scores of cognitive biomarkers. Ieee Access, 11, 123173-123193.
5. Cowen, A. S., & Keltner, D. (2021). Semantic space theory: A computational approach to emotion. Trends in Cognitive Sciences, 25(2), 124-136.
6. Dhanya, V. G., Subeesh, A., Kushwaha, N. L., Vishwakarma, D. K., Kumar, T. N., Ritika, G., & Singh, A. N. (2022). Deep learning based computer vision approaches for smart agricultural applications. Artificial Intelligence in Agriculture, 6, 211-229.
7. Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large language models for mental health applications: systematic review. JMIR mental health, 11(1), e57400.
8. Hung, L. P., & Alias, S. (2023). Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection. Journal of Advanced Computational Intelligence and Intelligent Informatics, 27(1), 84-95.
9. Kalateh, S., Estrada-Jimenez, L. A., Nikghadam-Hojjati, S., & Barata, J. (2024). A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges. IEEE Access, 12, 103976-104019.
10. Kišjuhas, A. (2024). What holds society together? Emotions, social ties, and group solidarity in leisure interaction rituals. Leisure Studies, 43(3), 363-377.
11. Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Systems with Applications, 244, 122778.
12. Mahfoudi, M. A., Meyer, A., Gaudin, T., Buendia, A., & Bouakaz, S. (2022). Emotion expression in human body posture and movement: A survey on intelligible motion factors, quantification and validation. IEEE Transactions on Affective Computing, 14(4), 2697-2721.
13. Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. Social network analysis and mining, 11(1), 81.
14. Tapeh, A. T. G., & Naser, M. Z. (2023). Artificial intelligence, machine learning, and deep learning in structural engineering: a scientometrics review of trends and best practices. Archives of Computational Methods in Engineering, 30(1), 115-159.
15. Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780.