Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

Comparative Analysis of Machine Learning Models for Emotion Prediction in Reviews of Large Language Model-Powered Applications

Semmy Wellem Taju [1*] George Morris William Tangka [1] Ibrena Reghuella Chrisanti [1]

[1] Universitas Klabat, Sulawesi Utara, Indonesia. Email: semmy@unklab.ac.id; gtangka@unklab.ac.id; ibrena@unklab.ac.id

ARTICLE HISTORY

Submitted	:	June 15, 2026
Reviewed	:	June 30, 2026
Revised	:	July 10, 2026
Accepted	:	August 01, 2026
Published	:	August 10, 2026

Conflict of Interest Statement:
The author(s) declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ABSTRACT

Purpose: This study compares the performance of conventional machine learning algorithms for multiclass emotion prediction in reviews of Large Language Model (LLM)-powered applications using a unified TF-IDF representation.

Research Method: User reviews were collected from the Google Play Store for eight popular LLM applications through web scraping. After text preprocessing, reviews were labeled into five emotion categories (Happy, Sadness, Anger, Fear, and Surprise). Ten machine learning algorithms, namely Perceptron, KNN, Naïve Bayes, Logistic Regression, MLP, SVM (Linear and RBF), Random Forest, XGBoost, and LightGBM, were evaluated using Accuracy, Balanced Accuracy, Precision, F1-score, MCC, and ROC-AUC.

Results and Discussion: XGBoost achieved the best performance, with 87.30% Accuracy, 77.84% Balanced Accuracy, 64.28% F1-score, and 0.5626 MCC. ROC-AUC values between 0.91 and 0.96 demonstrate excellent discrimination across all emotion classes. The superior performance of XGBoost indicates that gradient boosting effectively captures informative patterns from sparse TF-IDF features.

Implications: The findings provide practical guidance for selecting efficient machine learning models for emotion analysis of LLM application reviews and support AI application improvement through user emotion understanding.

Originality: This study presents a comprehensive benchmark of ten conventional machine learning algorithms using a large-scale multi-application LLM review dataset and a unified five-emotion classification framework.

Keywords: *Emotion prediction; Large language models; Machine learning; Text classification; TF-IDF.*

1. Introduction

The rapid advancement of Artificial Intelligence (AI), particularly Large Language Models (LLMs), has fundamentally transformed human–computer interaction by enabling users to perform a wide range of cognitive tasks, including information retrieval, content generation, programming assistance, language translation, and decision support. Consequently, LLM-powered applications such as ChatGPT (Wu et al., 2023), Gemini (Saeidnia, 2023), Claude (Liu et al., 2026), DeepSeek (Zhang et al., 2025), Grok (Adinath & IS, 2025), Microsoft Copilot (Stratton, 2024), Kimi (Gibney, 2025), and Qwen (Ahmed et al., 2025) have experienced substantial growth in global adoption. As their popularity continues to increase, digital

Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

application platforms such as the Google Play Store have accumulated millions of user reviews reflecting users' experiences, perceptions, and emotional responses toward application quality, usability, response accuracy, reliability, and overall user satisfaction. These reviews provide an important source of user-generated content for evaluating AI-powered services.

Previous studies have predominantly analyzed user reviews using sentiment analysis, which classifies opinions into positive, negative, or neutral categories (Cui et al., 2023; Mäntylä et al., 2018; Naseem et al., 2021; Nanli et al., 2012; Yue et al., 2019). Although sentiment analysis effectively summarizes general opinion polarity, it cannot adequately capture the diversity and intensity of users' emotional experiences. Emotion prediction offers a more fine-grained perspective by identifying discrete emotional states such as happiness, sadness, anger, fear, and surprise, thereby providing deeper insights into users' affective reactions toward AI systems (Chatterjee et al., 2019; Deng & Ren, 2021). Such information enables developers to better understand the underlying causes of user satisfaction and dissatisfaction, facilitating more effective improvements in AI application quality and user experience (Fu et al., 2013; Zhai et al., 2022).

Recent advances in text-based emotion recognition have led to the adoption of numerous machine learning algorithms, including Naïve Bayes, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, and LightGBM (Acheampong et al., 2020; Al Maruf et al., 2024; Ali et al., 2023; Berrar, 2018; Deng & Ren, 2021; Jakkula, 2006; Ke et al., 2017; Liu et al., 2012; Nusinovici et al., 2020; Zhang, 2016). These algorithms represent different learning paradigms, including probabilistic, linear, distance-based, ensemble, boosting, and neural network approaches. Their predictive capabilities vary substantially because of differences in model assumptions, decision boundaries, and their ability to learn from high-dimensional sparse textual features. In particular, TF-IDF remains one of the most widely adopted feature extraction techniques because it produces computationally efficient sparse vector representations while preserving discriminative lexical information for conventional machine learning algorithms (Al-Obaydy et al., 2022; Chen et al., 2021; Christian et al., 2016). Consequently, the interaction between TF-IDF representations and different classification algorithms warrants systematic investigation.

Despite the growing body of research on sentiment analysis and emotion recognition (Alam et al., 2025; Sharma et al., 2025; Tan et al., 2023), existing studies generally focus on sentiment polarity or evaluate only a limited number of classification algorithms (Vinodhini & Chandrasekaran, 2016). Furthermore, user reviews of LLM-powered applications exhibit linguistic characteristics that differ from those of conventional mobile applications because users frequently evaluate reasoning capability, contextual understanding, factual accuracy, privacy, safety, and response reliability. These distinctive characteristics may influence the effectiveness of different machine learning algorithms, yet comprehensive benchmarking studies within this emerging application domain remain scarce. Therefore, a systematic comparison is required to identify the most suitable conventional machine learning model for multiclass emotion prediction in LLM-powered application reviews.

To address this gap, this study collected user reviews from eight popular LLM-powered applications through Google Play Store web scraping (Chen et al., 2021; Latif et al., 2019). After text preprocessing, reviews were represented using TF-IDF and classified into five emotion categories: Happy, Sadness, Anger, Fear, and Surprise. The resulting feature vectors were used to evaluate ten

Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

machine learning algorithms, namely Perceptron, KNN, Naïve Bayes, Logistic Regression, Multi-Layer Perceptron (MLP), SVM with Linear and RBF kernels, Random Forest, XGBoost, and LightGBM. Model performance was assessed using Accuracy, Balanced Accuracy, Sensitivity, Specificity, Precision, F1-score, Matthews Correlation Coefficient (MCC), and ROC-AUC to provide a comprehensive evaluation of classification performance.

The contributions of this study are threefold. First, it constructs a large-scale dataset of user reviews collected from multiple LLM-powered applications available on the Google Play Store. Second, it develops a unified multiclass emotion prediction framework based on five discrete emotion categories using a consistent TF-IDF feature representation. Third, it provides a comprehensive benchmark of ten conventional machine learning algorithms using standardized evaluation metrics, thereby offering practical guidance for selecting computationally efficient emotion classification models and establishing a benchmark for future studies on emotion analysis in LLM-powered applications.

The remainder of this paper is organized as follows. Section 2 provides literature review. Section 3 presents research method. Section 4 presents discussion. Section 5 wrap up the paper.

2. Literature Review and Hypothesis Development

2.1 Emotion Prediction in User-Generated Reviews

User-generated reviews have become one of the most valuable sources of information for understanding users' experiences, perceptions, and behavioral intentions toward digital products and services. The rapid growth of online platforms has enabled users to express their opinions through textual reviews, providing organizations with abundant unstructured data that can be utilized to evaluate product quality, identify user concerns, and support decision-making processes (Fu et al., 2013; Zhai et al., 2022). Consequently, analyzing user-generated reviews has become an important research topic in natural language processing (NLP), text mining, and artificial intelligence.

Most previous studies have focused on sentiment analysis, which categorizes textual opinions into positive, negative, or neutral polarity (Cui et al., 2023; Mäntylä et al., 2018; Tan et al., 2023). Although sentiment analysis is effective for identifying general opinion trends, it provides only coarse-grained information regarding users' affective responses. Human emotions are inherently multidimensional, and reviews expressing the same sentiment polarity may represent substantially different emotional states. For example, two negative reviews may reflect disappointment, frustration, or anxiety, each implying different underlying user concerns and requiring different improvement strategies. Therefore, sentiment polarity alone is often insufficient for understanding the complexity of user experiences.

To address this limitation, recent studies have increasingly adopted emotion prediction, which aims to identify discrete emotional categories expressed in textual data, such as happiness, sadness, anger, fear, and surprise (Acheampong et al., 2020; Al Maruf et al., 2024; Chatterjee et al., 2019; Deng & Ren, 2021). Compared with sentiment analysis, emotion prediction provides a more fine-grained representation of users' psychological responses because it captures both the type and intensity of emotional expressions. This richer emotional information enables researchers and practitioners to better understand the motivations underlying user satisfaction, dissatisfaction, trust, anxiety, and engagement, thereby facilitating more targeted service improvements and user-centered system development.

The importance of emotion prediction becomes even more pronounced in the context of Large Language Model (LLM)-powered applications. Unlike conventional mobile applications, users of LLM

Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

systems evaluate not only interface usability and system performance but also the quality of generated responses, reasoning capability, contextual understanding, factual accuracy, privacy protection, safety, and reliability. These unique interaction characteristics often evoke complex emotional responses that cannot be adequately represented using simple positive–negative sentiment classification. Consequently, multiclass emotion prediction provides a more appropriate analytical framework for understanding users' affective experiences when interacting with generative AI systems.

Despite the increasing attention devoted to emotion recognition, empirical studies investigating emotion prediction specifically in reviews of LLM-powered applications remain limited. Existing research has predominantly examined sentiment analysis or focused on general social media and product review datasets, leaving the rapidly growing domain of LLM application reviews largely unexplored. This gap highlights the need for systematic research that investigates emotion prediction models capable of accurately identifying diverse emotional responses within user reviews of modern AI-powered applications.

2.2 Machine Learning for Text-based Emotion Classification

Machine learning has become one of the most widely adopted approaches for text-based emotion classification because of its ability to automatically learn discriminative patterns from textual data. Unlike rule-based methods, machine learning algorithms can model complex relationships between textual features and emotion categories, making them suitable for large-scale text mining applications (Acheampong et al., 2020; Deng & Ren, 2021). However, their predictive performance depends not only on the quality of feature representation but also on the underlying learning mechanism of each algorithm. Since textual features generated by methods such as Term Frequency–Inverse Document Frequency (TF-IDF) are typically high-dimensional and sparse, different machine learning paradigms exhibit different strengths and limitations when processing such representations.

Probabilistic classifiers, particularly Naïve Bayes, estimate the probability of each class based on Bayes' theorem while assuming conditional independence among input features (Berrar, 2018). This assumption enables efficient computation and fast model training but often oversimplifies the relationships among textual features. Consequently, Naïve Bayes may experience performance degradation when important emotional expressions depend on interactions among multiple words rather than independent term occurrences.

Linear classifiers, including Logistic Regression and Support Vector Machine (SVM), have consistently demonstrated strong performance in text classification tasks because they are well suited for sparse and high-dimensional feature spaces (Jakkula, 2006; Nusinovici et al., 2020). These algorithms construct linear decision boundaries that maximize class separability while reducing the risk of overfitting. Their computational efficiency and robustness make them widely adopted baselines for document classification using TF-IDF representations.

Distance-based methods such as K-Nearest Neighbors (KNN) classify documents according to the similarity between feature vectors (Zhang, 2016). Although KNN is conceptually simple and requires minimal model assumptions, its effectiveness often decreases in high-dimensional text spaces due to the curse of dimensionality, where distance measurements become less discriminative. Consequently, KNN generally performs less effectively than linear classifiers on sparse textual datasets.

Neural network models, such as the Multi-Layer Perceptron (MLP), are capable of learning nonlinear relationships among textual features through multiple hidden layers (Chatterjee et al., 2019). While MLP offers greater representational flexibility than linear models, its performance depends heavily on the availability of sufficient training data and appropriate hyperparameter configuration. Without

Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

adequate optimization, neural networks may not fully exploit sparse lexical representations generated by TF-IDF.

Ensemble learning methods improve predictive performance by combining multiple decision trees. Random Forest employs a bagging strategy that constructs independent decision trees using bootstrap sampling and feature randomness, thereby reducing model variance and improving generalization (Liu et al., 2012). In contrast, gradient boosting algorithms, including XGBoost and LightGBM, sequentially construct decision trees in which each new tree focuses on correcting the prediction errors made by previous trees (Ali et al., 2023; Bentéjac et al., 2019; Ke et al., 2017). This iterative learning process enables boosting algorithms to capture complex nonlinear feature interactions and improve classification accuracy, particularly in high-dimensional datasets containing sparse yet informative lexical features.

Given these fundamental differences in learning mechanisms, machine learning algorithms are expected to exhibit different predictive capabilities when applied to TF-IDF-based emotion classification. Therefore, a comprehensive comparison across probabilistic, linear, distance-based, neural network, bagging, and boosting approaches is essential to identify the most appropriate classifier for multiclass emotion prediction in reviews of LLM-powered applications.

2.3 TF-IDF Representation for Sparse Text Classification

Feature representation plays a fundamental role in text classification because the predictive performance of machine learning algorithms largely depends on how textual information is transformed into numerical features. Since machine learning models cannot directly process unstructured textual data, an effective feature extraction technique is required to preserve discriminative information while maintaining computational efficiency. Among various text representation methods, Term Frequency–Inverse Document Frequency (TF-IDF) remains one of the most widely adopted approaches because of its simplicity, interpretability, and competitive performance in document classification tasks (Al-Obaydy et al., 2022; Christian et al., 2016).

TF-IDF assigns a numerical weight to each term according to its frequency within a document and its rarity across the entire corpus. Terms that frequently appear in a particular document but rarely occur in other documents receive higher weights because they provide greater discriminative power, whereas common words appearing in most documents receive lower weights (Christian et al., 2016). Consequently, TF-IDF emphasizes informative lexical features while reducing the influence of less meaningful terms, producing a document-term matrix that effectively represents textual content for supervised learning algorithms.

Despite its effectiveness, TF-IDF generates high-dimensional and sparse feature representations, in which each document contains only a small proportion of the vocabulary. This sparsity has important implications for classifier performance because different machine learning algorithms exploit sparse representations in different ways. Linear models such as Logistic Regression and Support Vector Machine are generally well suited to sparse feature spaces because they estimate decision boundaries directly from informative weighted features. Probabilistic methods such as Naïve Bayes benefit from computational efficiency but may be constrained by their conditional independence assumption. In contrast, ensemble boosting methods such as XGBoost and LightGBM are capable of identifying complex nonlinear interactions among informative features by sequentially correcting classification errors, making them particularly effective when discriminative terms are unevenly distributed across documents (Ali et al., 2023; Bentéjac et al., 2019; Ke et al., 2017).

Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

Although recent transformer-based language models provide contextual semantic embeddings, TF-IDF continues to offer several practical advantages for conventional machine learning research. It requires substantially lower computational resources, provides highly interpretable feature weights, and enables fair comparisons across different classification algorithms because all models are trained using an identical feature representation (Chen et al., 2021). For benchmarking studies involving multiple conventional classifiers, TF-IDF therefore remains an appropriate and widely accepted baseline for evaluating algorithmic behavior under consistent experimental conditions.

Given these characteristics, this study employs TF-IDF as a unified feature extraction method to ensure that differences in predictive performance primarily reflect the learning mechanisms of the evaluated machine learning algorithms rather than differences in feature representation. This approach enables a systematic and objective comparison of classification performance for multiclass emotion prediction in reviews of LLM-powered applications.

3. Research Method

3.1. Research Framework

This study uses a *text mining approach* to predict emotions in user reviews of a Large Language Model (LLM)-based AI Chat application. The research stages consist of data collection, text preprocessing, feature extraction using TF-IDF, machine learning model training, and model performance evaluation. The research flow is shown in Figure 1.

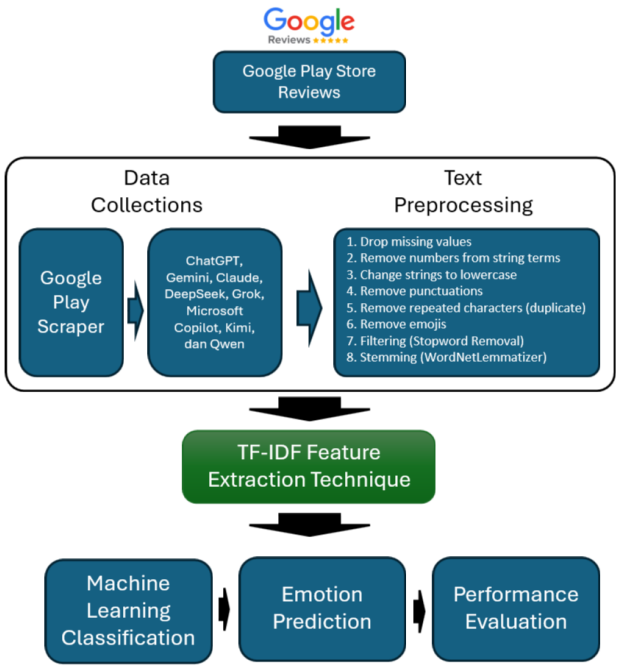

Figure 1. Research Framework

3.1.1. Data Collection

In this study, the dataset was obtained through a web scraping process using the Google Play Scraper library called *google-play-scraper* [36] in the Python programming language. Data was taken from user reviews of LLM-based applications available on the Google Play Store, such as ChatGPT, Gemini, Claude, DeepSeek, Grok, Microsoft Copilot, Kimi and Qwen. The information collected includes: *reviewId,*

Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

content, score, thumbsUpCount, reviewCreatedVersion, at, replyContent, repliedAt and appVersion. Data retrieval was carried out automatically using language parameters (English), country (United States), number of reviews according to research needs, and certain rating categories if necessary. Table 1 shows the distribution of the dataset used in this study. The dataset consists of 32563 user reviews collected from eight LLM-based AI Chat applications available on the Google Play Store, namely ChatGPT, Claude, Microsoft Copilot, DeepSeek, Gemini, Grok, Kimi, and Qwen. Most of the review data from each downloaded application has 5000 reviews, while Kimi and Qwen have a smaller number of reviews due to the limited data available at the time of the web scraping process.

Table 1.

Summary of the Collected Review Dataset Across LLM Applications

Application	Original Reviews	After Pre-Processing	Coverage (%)
ChatGPT	5,000 Reviews	2,413	48.26
Claude	5,000 Reviews	2,123	42.46
Microsoft Copilot	5,000 Reviews	2,365	47.3
DeepSeek	5,000 Reviews	1,935	38.7
Gemini	5,000 Reviews	2,754	55.08
Grok	5,000 Reviews	2,048	40.96
Kimi	2,207 Reviews	693	31.4
QWen	356 Reviews	31	8.71

After text preprocessing, each review was classified into five emotion categories: Happy, Sadness, Anger, Fear, and Surprise using a keyword-based emotion labeling approach. The keyword list was compiled based on common emotional expressions used in English-language reviews. Each review was examined for the presence of words or phrases representing one of the emotion categories.

If a review contained a matching keyword, it was assigned to the corresponding emotion category. The labeling process showed that the Surprise category had the largest number of reviews, with 3,996 unique reviews (27.82%), followed by Happy with 3,591 reviews (25.01%), and Sadness with 3,200 reviews (22.28%). Meanwhile, the Anger category consisted of 2,274 reviews (15.83%), while the Fear category had the smallest number of reviews, with 1,301 reviews (9.06%). Overall, 14,362 unique reviews were obtained, which were used as the dataset for the training and evaluation of the Machine Learning model. The dataset was divided into training and testing sets using an 80:20 ratio, where 80% of the data were used for model training and the remaining 20% were used for performance evaluation.

3.1.2. Emotion Labeling

This study formulates emotion prediction as a multiclass classification problem with five emotion categories used such as Happy category is a review that shows satisfaction, joy, or appreciation for the application. Sadness category is a review that shows disappointment, loss, or sadness. Anger category is a review that contains anger, frustration, or strong criticism. Fear category is a review that shows worry, anxiety, or fear regarding the use of the application. Surprise category is a review that shows surprise, admiration, or surprise regarding the features or results of the application. Each review is

Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

labeled with an emotion before the model training process so that it can be used as multiclass classification data.

3.1.3. Text Preprocessing

The text preprocessing is carried out to improve the quality of text data before the feature extraction and classification process. This process aims to reduce noise, standardize text format, and remove irrelevant information so as to improve the performance of the classification model. In this study, the preprocessing stages consist of case folding, cleaning, tokenization, stopword removal, and lemmatization. The first stage is case folding, which is to clean letters from capital letters to lowercase and also remove numbers and existing punctuation. Some steps carried out are (1) Drop missing values, (2) Remove numbers from string terms, (3) Change strings to lowercase, (4) Remove punctuations (annotation removal), (5) Remove repeated characters (*duplicate*) characters from a string and (6) Remove emojis. By following these initial stages, all characters become lowercase and can also eliminate the difference between capital and lowercase letters. Next, the cleaning process is carried out to remove elements that do not contribute to the classification process, such as URLs, HTML tags, email addresses, numbers, punctuation, emojis and special characters. This stage aims to produce cleaner and more consistent text. After the cleaning process, tokenization is performed to break each sentence into a collection of words (tokens). These tokens are then processed using stopword removal to remove common words that have a high frequency of occurrence but do not provide significant information to the sentence's meaning, such as *the, is, are, a, an*, and *of*. The Natural Language Toolkit (NLTK) library is used to obtain a collection of stopwords. The final stage is *lemmatization*, which converts each word to its base form (*lemma*) based on linguistic rules so that variations of word forms, such as running, runs, and ran, are represented as the base word run. The NLTK library is also utilized in this process with the WordNetLemmatizer function used to transform each word. This process helps reduce the dimensionality of features while increasing the consistency of word representation. The result of all text preprocessing stages is a collection of cleaned and normalized text documents. The preprocessed dataset is then used as input for the feature extraction stage using the TF-IDF method before the classification process using various Machine Learning algorithms.

3.2. Feature Extraction

Text feature representation is built using the Term Frequency-Inverse Document Frequency (TF-IDF) method. In this study, researchers hope this method can transform unstructured text data into a numeric representation that can be processed by Machine Learning algorithms. TF-IDF works by assigning weights to each word based on its level of importance in a document relative to the entire document collection (corpus). Words that appear frequently in a document but rarely appear in other documents will receive a higher weight because they are considered to have better discriminatory ability. Conversely, words that appear in almost all documents will receive a lower weight because they have a small contribution in distinguishing document classes. This method is divided into two terms, where TF measures how often a word appears in a document or means the more frequently the word appears, the higher the TF value. Meanwhile, IDF measures how important or rare a word is in the entire document collection (*corpus*). If a word appears in almost all documents, then the word is not more important. IDF will give a low weight to common words and a high weight to unique words. Thus, TF-IDF is able to produce a more informative feature representation than using term frequency alone. Mathematically, the TF-IDF weight is obtained from the multiplication of the Term Frequency (TF) and

Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

Inverse Document Frequency (IDF) values. The TF value indicates the frequency of a word's appearance in a document, while the IDF value measures the word's rarity across all documents in the corpus. The TF-IDF weight is calculated using Equation (1), while the TF value is calculated using Equation (2) and the IDF value is calculated using Equation (3).

Term Frequency–Inverse Document Frequency (TF-IDF):

$$TF\text{-}IDF = TF \times IDF \quad (1)$$

Term Frequency (TF):

$$TF(t,d) = \frac{Number\ of\ occurrences\ of\ term\ t\ in\ document\ d}{Total\ number\ of\ words\ in\ document\ d} \quad (2)$$

Inverse Document Frequency (IDF):

$$IDF(t,D) = log\left(\frac{N}{DF(t)}\right) \quad (3)$$

As shown in the TF-IDF formulation (Equation 1, Equation 2 and Equation 3), *t* represents the term (*word*) whose frequency or weight is being measured, while *d* represents the document in which the term appears. *D* represents the collection of all documents (corpus), and *N* is the total number of documents within the corpus. *DF(t)* or Document Frequency represents the number of documents that contain the term *t*. Finally, *log* refers to the *logarithmic function*, which is applied to reduce the scale of the resulting values. The result of the TF-IDF transformation is a high-dimensional feature matrix (document-term matrix), where each row represents a document and each column represents a term along with its TF-IDF weight. This matrix is then used as input for all Machine Learning algorithms compared in this study, namely Perceptron, KNN, MLP, Naïve Bayes, Support Vector Machine (Linear and RBF), Random Forest, Logistic Regression, XGBoost, and LightGBM. The use of the same feature representation in this study across all algorithms aims to ensure that the differences in performance obtained come from the characteristics of each classification algorithm, so that the comparison results can be carried out fairly (fair comparison) and objectively.

3.2.1. Machine Learning Classification

A comparative analysis of several Machine Learning algorithms was conducted to identify the most effective model in predicting emotions in LLM-based AI Chat application reviews. The feature representation generated using the TF-IDF method was then used as input for each classification algorithm. The algorithms evaluated included Naïve Bayes, K-Nearest Neighbor (KNN), Multi-Layer Perceptron (MLP), Perceptron, Support Vector Machine (SVM) with Linear and Radial Basis Function (RBF) kernels, Random Forest, XGBoost and LightGBM. The selection of these algorithms was based on their characteristics that represent various classification approaches, ranging from probabilistic, distance-based, linear, ensemble learning, to gradient boosting models. To ensure that the comparison was conducted fairly and produced reproducible experiments, each algorithm was implemented using

Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

the same parameter configuration throughout the entire experimental process. Most parameters use the default values provided by the Scikit-learn, XGBoost, and LightGBM libraries, while some parameters are adjusted, such as the number of neighbors in KNN, the number of estimators in Random Forest, XGBoost, and LightGBM, as well as certain parameters in Perceptron, MLP, and SVM to maintain the consistency of the training process. Details of the parameter configuration for each algorithm are presented in Table 2.

Table 2.

Hyperparameter Settings of the Machine Learning Models Used for Emotion Prediction

No.	Algorithms	Parameters
1	Naive Bayes	default
2	KNN	n_neighbors=300
3	LightGBM	n_estimators=100, random_state=0
4	MLP	hidden_layer_sizes=(100,), max_iter=500, random_state=0
5	Perceptron	max_iter=100, eta0=0.9, random_state=0
6	XGBoost	n_estimators=100, random_state=0
7	Random Forest	n_estimators=100, max_depth=10, random_state=0
8	SVM (Linear)	kernel='linear', gamma=0.1, C=1.0, probability=False
9	SVM (RBF)	kernel='rbf', gamma=0.1, C=1.0, probability=False
10	Logistic Regression	random_state=0, max_iter=1000

Table 2 shows the parameter configurations used in each Machine Learning algorithm. Naïve Bayes was implemented using default parameters. KNN used *n_neighbors* = 300 to determine the number of nearest neighbors in the classification process. MLP was configured with *hidden_layer_sizes* = 100, *max_iter* = 500, and *random_state* = 0, while Perceptron used *max_iter* = 100, *eta0* = 0.9, and *random_state* = 0. Random Forest, XGBoost and LightGBM each used 100 *estimators*, while Random Forest was also limited to *max_depth* = 10 to reduce model complexity. In the SVM algorithm, the Linear kernel used the parameter *C* = 1.0, while the RBF kernel used *C* = 1.0 and *gamma* = 0.1. In addition, all models that supported the *random_state* setting used the value 0 to ensure reproducible experimental results.

Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

3.2.2. Performance Evaluation

Model performance was evaluated using several classification metrics calculated based on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values obtained from the confusion matrix. The evaluation metrics used include Sensitivity (Recall), Specificity, Accuracy, Balanced Accuracy, Precision, F1-Score, and Matthews Correlation Coefficient (MCC). Sensitivity (Recall) measures the model's ability to correctly identify positive data. Specificity measures the model's ability to correctly identify negative data. Accuracy indicates the proportion of all predictions that are correctly classified. Balanced Accuracy is the average of Sensitivity and Specificity, making it more suitable for use in imbalanced datasets. Precision measures the proportion of correct positive predictions. F1-Score is the harmonic mean between Precision and Recall. Matthews Correlation Coefficient (MCC) is an evaluation metric that considers all elements in the confusion matrix, making it more representative for evaluating models on imbalanced datasets. MCC values range from -1 to 1, with MCC = 1 indicating perfect classification. MCC = 0 indicates performance equivalent to random prediction. MCC = -1 indicates a completely incorrect classification. Since this study is a multi-class classification consisting of five emotion categories (Happy, Sadness, Anger, Fear, and Surprise), the evaluation metrics are calculated for each class and averaged, so that each emotion category contributes equally to the overall evaluation result.

$$Sensitivity\ (Sens) = \frac{TP}{TP+FN} \quad (4)$$

$$Specificity\ (Spec) = \frac{TN}{TN+FP} \quad (5)$$

$$Accuracy\ (Acc) = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Balanced\ Accuracy\ (BalAcc) = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (7)$$

$$Precision\ (Prec) = \frac{TP}{TP+FP} \quad (8)$$

$$F1 - Score\ (FS) = \frac{2TP}{2TP+FP+FN} \quad (9)$$

$$Matthews\ Correlation\ Coefficient\ (MCC) = \frac{TPxTN-FPxFN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (10)$$

Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

4. Results and Discussion

4.1. Distribution of Review Categories Across LLM Applications

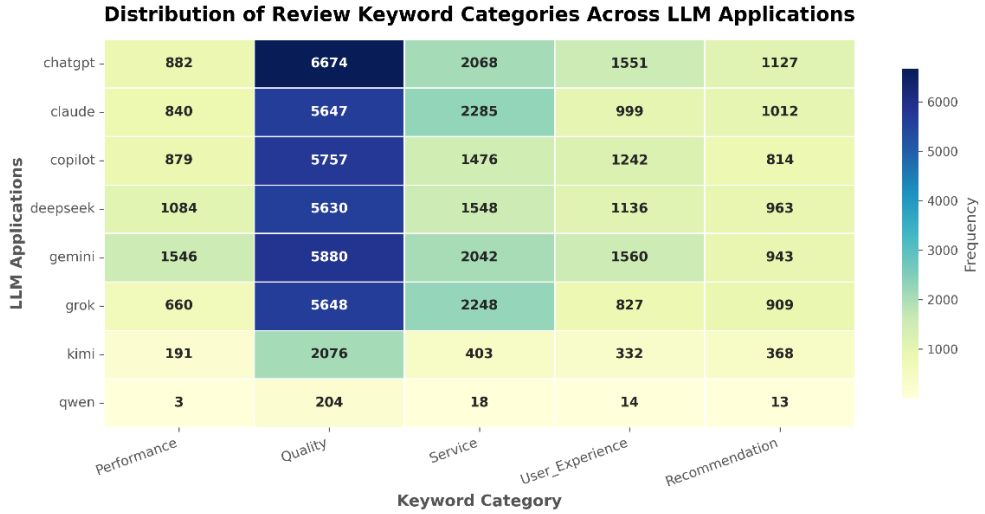

Figure 2.

Comparative distribution of review categories Large Language Model (LLM) applications

Figure 2 shows the comparative distribution of review keyword categories for eight LLM-based AI Chat applications such as ChatGPT, Claude, Microsoft Copilot, DeepSeek, Gemini, Grok, Kimi, and Qwen. Overall, the Quality category dominated discussions across all applications, with ChatGPT having the highest number of keywords (6,674 keywords), followed by Gemini (5,880 keywords), Copilot (5,757 keywords), Grok (5,648 keywords), Claude (5,647 keywords), and DeepSeek (5,630 keywords). These results indicate that users pay more attention to the quality of the application's output, such as accuracy, relevance, and reliability of responses. The Service category was the most discussed aspect after Quality, reflecting users' attention to the application's services, features, and functionality.

Meanwhile, the Performance and User Experience categories also received a relatively high number of keywords, indicating that response speed, system stability, ease of use, and interface comfort also influence user experience. Conversely, the Recommendation category had the lowest frequency across all apps, indicating that users tended to share their experiences and assessments of the app's quality and functionality rather than explicitly recommending it to others. Overall, these results indicate that response quality is a key factor influencing users' perceptions and evaluations of LLM-based apps.

4.1.1. Performance Comparison of Machine Learning Models

Table 3 presents a comparison of the performance of nine Machine Learning algorithms for emotion prediction using TF-IDF feature representation. Based on the test results, XGBoost provides the best performance with an Accuracy of 87.30%, Balanced Accuracy of 77.84%, Sensitivity of 63.84%, Specificity of 91.82%, Precision of 65.04%, F1-Score of 64.28%, and MCC of 0.5626. This performance is followed by SVM Linear and SVM RBF which each obtained an Accuracy of 87.28% with an MCC of 0.5564, while Logistic Regression also showed competitive results with an Accuracy of 87.62% and an MCC of 0.5623.

Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

In contrast, Naïve Bayes produced the lowest performance with an Accuracy of 66.18%, an F1-Score of 15.88%, and a negative MCC (-0.0064), indicating the model's low ability to distinguish emotion categories. In general, these results indicate that gradient boosting-based algorithms and linear models, such as XGBoost, SVM, and Logistic Regression, are more effective in handling high-dimensional and sparse TF-IDF feature representations, thus producing better classification performance than other algorithms.

Table 3.

Performance Evaluation Results of Machine Learning Models for Emotion Classification

Algorithms	Acc	BalAcc	Sens	Spec	Prec	FS	MCC
Naïve Bayes	66.18%	49.1%	18.3%	79.92%	20.08%	15.88%	-0.0064
KNN	81.8%	65.76%	43.82%	87.74%	67.44%	39.3%	0.3438
Logistic Regression	87.62%	77.24%	62.52%	91.94%	66.68%	63.36%	0.5623
Perceptron	82.92%	71.46%	53.82%	89.16%	54.26%	52.86%	0.4294
MLP	82.04%	69.48%	50.46%	88.5%	50.94%	50.64%	0.392
SVM Linear	87.28%	77.3%	62.8%	91.76%	65.02%	63.46%	0.5564
SVM RBF	87.28%	77.3%	62.8%	91.76%	65.02%	63.46%	0.5564
Random Forest	81.1%	64.04%	41.02%	87.08%	41.22%	35.9%	0.2892
XGBoost	**87.3%**	**77.84%**	**63.84%**	**91.82%**	**65.04%**	**64.28%**	**0.5626**
LightGBM	86.26%	76.02%	60.88%	91.22%	61.64%	61.18%	0.5248

Table 4 presents the best model performance for each emotion category. The evaluation results show that the model has the best ability to identify the Happy emotion, as indicated by a Sensitivity value of 81.60%, Precision of 79.90%, F1-Score of 80.70%, Balanced Accuracy of 87.40%, and MCC of 0.742. Conversely, the Fear category obtained the lowest Sensitivity value (40.60%) despite having the highest Specificity (96.10%), indicating that the model is quite good at distinguishing reviews that do not belong to the Fear category, but still has difficulty in recognizing reviews that actually contain this emotion.

The Sadness and Surprise categories showed relatively balanced performance with F1-Score values of 68.70% and 69.00%, respectively, while the Anger category obtained an F1-Score of 57.80%. Overall, the high Specificity value in all emotion categories (86.70%-96.10%) indicates that the model is able to reduce classification errors against the negative class, while variations in Sensitivity values indicate that the model's ability to recognize each emotion category is still influenced by linguistic characteristics and data distribution in each class.

Table 4.

Class-wise Performance of the Best Machine Learning Model for Emotion Prediction

Emotion Class	Acc	BalAcc	Sens	Spec	Prec	FS	MCC
Anger	86.90%	74.60%	56.70%	92.60%	59.00%	57.80%	0.501
Fear	**91.10**%	68.40%	40.60%	**96.10**%	51.00%	45.20%	0.407
Happy	90.30%	**87.40**%	**81.60**%	93.10%	**79.90**%	**80.70**%	**0.742**
Sadness	85.90%	80.00%	69.50%	90.60%	68.00%	68.70%	0.597

Advances in Human Resource Management Research
Volume 4, Issue 3 (2026). 404-422
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008
OPEN ACCESS
e-ISSN: 2985-7570

Surprise	82.30%	78.80%	70.80%	86.70%	67.30%	69.00%	0.566

4.1.2. ROC Curves

Figure 3 shows the Receiver Operating Characteristic (ROC) curves for each emotion category generated by the XGBoost model. It can be seen that all ROC curves are well above the diagonal line (random classifier line), indicating that the model has excellent discriminatory ability in distinguishing each emotion category compared to random classification. Based on the Area Under the Curve (AUC) value, the Happy category obtained the best performance with an AUC value of 0.96, indicating that the model is able to distinguish reviews containing the Happy emotion.

Meanwhile, the Sadness, Surprise, Anger, and Fear categories obtained AUC values of 0.93, 0.92, 0.91, and 0.91, respectively, which are also included in the excellent classification performance category (AUC > 0.90). The difference in AUC values between categories is relatively small, indicating that the model has consistent classification ability across all emotion classes. Overall, the ROC evaluation results confirm that the XGBoost model is able to provide good class separability in text-based emotion prediction tasks.

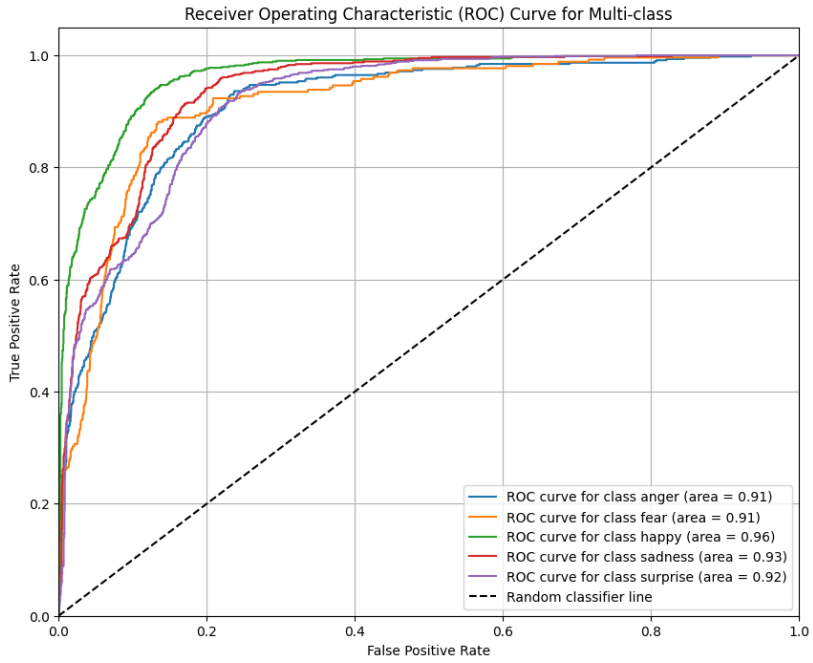

Figure 3.

Multi-class ROC curves of the XGBoost model for emotion prediction across five emotion categories

Advances in Human Resource Management Research
OPEN ACCESS
Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

4.1.3. Feature Importance of XGBoost Model

Figure 4 displays the 20 most important TF-IDF features identified by the XGBoost model based on the F-score value, which indicates the relative contribution of each feature to the model's decision-making process. Based on the results obtained, the word "*helpful*" has the highest level of importance, followed by "*privacy*", "*accurate*", "*love*", and "*dark*". The dominance of these words indicates that user perceptions of usability, privacy, response accuracy, and user experience are the main factors influencing emotion prediction in this LLM-based AI Chat application review. Furthermore, the presence of words such as "*wonderful*", "*incredible*", "*terrible*", "*garbage*", "*unreliable*", and "*dangerous*" indicates that the model also utilizes evaluative expressions representing positive and negative emotions in distinguishing emotion categories. Information regarding the importance of each of these features can be the basis for the feature engineering process and the development of more effective emotion prediction models in future research.

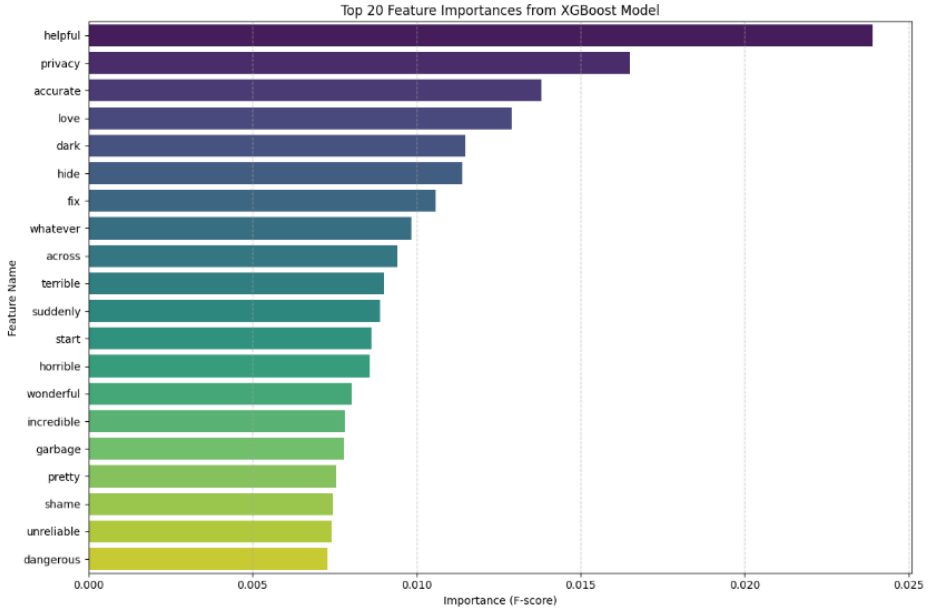

Figure 4.

Top 20 most important TF-IDF features identified by the XGBoost model for emotion prediction

5. Concluding Remarks and Recommendation

This study has analyzed nine Machine Learning algorithms for emotion prediction tasks in LLM-based AI Chat app reviews using TF-IDF feature representation. The research dataset was collected from user reviews of eight most popular LLM AI apps on the Google Play Store and classified into five emotion categories, namely Happy, Sadness, Anger, Fear, and Surprise. The experimental results show that XGBoost is the best performing model, which obtained an Accuracy of 87.30%, Balanced Accuracy of 77.84%, Precision of 65.04%, F1-Score of 64.28%, and MCC of 0.5626. In addition, the Receiver Operating Characteristic (ROC) analysis shows a high Area Under the Curve (AUC) value across all emotion categories (0.91-0.96), which indicates that the model has good discrimination ability in distinguishing each emotion class.

Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

Feature importance analysis also shows that words such as *helpful*, *privacy*, *accurate*, and *love* are the most contributing features in the emotion prediction process, thus providing insight into the aspects that users pay most attention to when evaluating LLM-based AI Chat applications. Overall, the results of this study indicate that gradient boosting-based algorithms, specifically XGBoost, are more effective than other Machine Learning algorithms in handling high-dimensional and sparse TF-IDF feature representations. It can be seen that these findings contribute to the selection of Machine Learning algorithms for the task of emotion prediction in LLM-based application reviews and can be a reference for application developers in understanding user perceptions and emotional responses more deeply. In future research, model performance can be improved through the application of hyperparameter tuning, the use of transformer-based feature extraction techniques such as BERT, RoBERTa, DeBERTa, or Sentence-BERT, and evaluation on larger and more diverse datasets to improve the model's generalization ability.

Statement of Use of Generative AI

During the preparation of this work, the author used generative artificial intelligence tools to support the scientific writing process. Grammarly was used to check grammar, refine writing style, and improve clarity in scientific writing. All interpretations, analyses, and conclusions presented in this study are the sole responsibility of the author.

Acknowledgment

The authors would like to express their sincere appreciation to all individuals and organizations who have directly or indirectly contributed to the completion of this research. The authors also acknowledge the developers of the open-source software and libraries used in this study, which greatly facilitated the processes of data collection, preprocessing, model development, and performance evaluation. Their valuable contributions have been instrumental in the successful completion of this work.

References

Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7), e12189. https://doi.org/10.1002/eng2.12189

Adinath, D., & IS, S. (2025). ChatGPT-5 and Grok-4: A comparative analysis of advancements in NLP. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5547358

Ahmed, I., Islam, S., Datta, P. P., Kabir, I., Chowdhury, M. N. U. R., & Haque, A. (2025). Qwen 2.5: A comprehensive review of the leading resource-efficient LLM with potential to surpass all competitors. https://doi.org/10.36227/techrxiv.174060306.65738406/v1

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. https://doi.org/10.1109/ACCESS.2024.3356357

Al-Obaydy, W. I., Hashim, H. A., Najm, Y. A., & Jalal, A. A. (2022). Document classification using term frequency-inverse document frequency and K-means clustering. *Indonesian Journal of Electrical Engineering and Computer Science*, *27*(3), 1517-1524.

Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

Alam, M. S., Mrida, M. S. H., & Rahman, M. A. (2025). Sentiment analysis in social media: How data science impacts public opinion knowledge integrates natural language processing (NLP) with artificial intelligence (AI). *American Journal of Scholarly Research and Innovation*, *4*(01), 63-100. https://doi.org/10.63125/r3sq6p80

Ali, Z. A., Abduljabbar, Z. H., Tahir, H. A., Sallow, A. B., & Almufti, S. M. (2023). eXtreme gradient boosting algorithm with machine learning: A review. *Academic Journal of Nawroz University*, *12*(2), 320-334. 10.25007/ajnu.v12n2a1612

Asniwati, A., & Latief, F. (2026). The Influence of Interpersonal Communication and Workplace Friendships on Employee Job Satisfaction in Public Organizations. *Advances in Human Resource Management Research*, *4*(2), 185-200. https://doi.org/10.60079/ahrmr.v4i2.748

Bahasoan, S., Asniwati, A., & Surianto, S. (2026). Synergy Between Training and Digital Capabilities: Transforming Competencies to Achieve Competitive Advantage for Retail Apparel SMEs. *Advances in Human Resource Management Research*, *4*(1), 89-111. https://doi.org/10.60079/ahrmr.v4i1.754

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). A comparative analysis of XGBoost. arXiv. https://arxiv.org/abs/1911.01914

Berrar, D. (2018). Bayes' theorem and naive Bayes classifier. In Encyclopedia of Bioinformatics and Computational Biology (pp. 403–412).

Chatterjee, A., Gupta, U., Chinnakotla, M. K., Srikanth, R., Galley, M., & Agrawal, P. (2019). Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, *93*, 309-317. https://doi.org/10.1016/j.chb.2018.12.029

Chen, Q., Chen, C., Hassan, S., Xing, Z., Xia, X., & Hassan, A. E. (2021). How should i improve the ui of my app? a study of user reviews of popular apps in the google play. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *30*(3), 1-38. https://doi.org/10.1145/3447808

Christian, H., Agus, M. P., & Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, *7*(4), 285-294.

Cui, J., Wang, Z., Ho, S. B., & Cambria, E. (2023). Survey on sentiment analysis: evolution of research methods and topics. *Artificial intelligence review*, *56*(8), 8469-8510. https://doi.org/10.1007/s10462-022-10386-z

Deng, J., & Ren, F. (2021). A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, *14*(1), 49-67. https://doi.org/10.1109/TAFFC.2021.3053275

Fakhri, M., & Silvianita, A. (2026). Servant Leadership and Employee Engagement Among Generation Z: Examining the Mediating Role of Mental Well-Being. *Advances in Human Resource Management Research*, *4*(2), 219-238. https://doi.org/10.60079/ahrmr.v4i2.772

Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J., & Sadeh, N. (2013, August). Why people hate your app: Making sense of user feedback in a mobile app store. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1276-1284).

Gibney, E. (2025). Chinese AI model Kimi K2 excites researchers. Nature, 643, 889.

Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

Hartati, H., Setiawan, A., Alyyasa-Gan, S. S., & Machmud, M. (2026). The Influence of Leadership and Motivation on Performance in High-Discipline Organizations: A Study of Brimob Personnel. *Advances in Human Resource Management Research*, *4*(1), 55-71. https://doi.org/10.60079/ahrmr.v4i1.749

Hidayat, N., Mojolou, D. N., Madli, F., Lada, S., & Loong, A. H. (2026). The Role of Trust and Knowledge Sharing in Encouraging Innovative Work Behavior: An Empirical Study of SMEs in Tarakan City. *Advances in Human Resource Management Research*, *4*(1), 1-13. https://doi.org/10.60079/ahrmr.v4i1.733

Jakkula, V. (2006). Tutorial on support vector machine (SVM). School of EECS, Washington State University.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Kosasih, K., Yesika, M., Lustina, F. A., Rusdiana, E., & Tansil, A. R. (2026). Artificial Intelligence in Hospital Human Resource Management: A Systematic Review and Bibliometric Analysis. *Advances in Human Resource Management Research*, *4*(2), 267-285. https://doi.org/10.60079/ahrmr.v4i2.776

Latif, R. M. A., Abdullah, M. T., Shah, S. U. A., Farhan, M., Ijaz, F., & Karim, A. (2019). Data scraping from Google Play Store and visualization of its content for analytics. In Proceedings of the 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) (pp. 1–8). IEEE.

Liu, J., Zhao, X., Shang, X., & Shen, Z. (2026). Dive into Claude Code: The design space of today's and future AI agent systems. arXiv. https://arxiv.org/abs/2604.14228

Liu, Y., Wang, Y., & Zhang, J. (2012, September). New machine learning algorithm: Random forest. In *International conference on information computing and applications* (pp. 246-252). Berlin, Heidelberg: Springer Berlin Heidelberg.

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer science review*, *27*, 16-32. https://doi.org/10.1016/j.cosrev.2017.10.002

Nanli, Z., Ping, Z., Weiguo, L., & Meng, C. (2012). Sentiment analysis: A literature review. In Proceedings of the International Symposium on Management of Technology (ISMOT) (pp. 572–576). IEEE.

Naseem, S., Mahmood, T., Asif, M., Rashid, J., Umair, M., & Shah, M. (2021). Survey on sentiment analysis of user reviews. In Proceedings of the 2021 International Conference on Innovative Computing (ICIC) (pp. 1–6). IEEE.

Nurinaya, N., & Marhumi, S. (2025). Leadership strategy evaluation and succession planning in preparing the company's future leaders. *Advances in Human Resource Management Research*, *3*(1), 1-14. https://doi.org/10.60079/ahrmr.v3i1.405

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., ... & Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, *122*, 56-69. https://doi.org/10.1016/j.jclinepi.2020.03.002

Volume 4, Issue 3 (2026). 404-422
e-ISSN: 2985-7570
DOI: https://doi.org/10.60079/ahrmr.v4i3.1008

Saeidnia, H. R. (2023). Welcome to the Gemini era: Google DeepMind and the information industry. Library Hi Tech News.

Sharma, N. A., Ali, A. S., & Kabir, M. A. (2025). A review of sentiment analysis: tasks, applications, and deep learning techniques. *International journal of data science and analytics*, *19*(3), 351-388. https://doi.org/10.1007/s41060-024-00594-x

Stratton, J. (2024). An introduction to Microsoft Copilot. In Copilot for Microsoft 365: Harness the power of generative AI in the Microsoft apps you use every day (pp. 19–35). Springer.

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, *13*(7), 4550. https://doi.org/10.3390/app13074550

Venkatakrishnan, S., Kaushik, A., & Verma, J. K. (2020). Sentiment analysis on google play store data using deep learning. In *Applications of Machine Learning* (pp. 15-30). Singapore: Springer Singapore.

Vinodhini, G., & Chandrasekaran, R. M. (2016). A comparative performance evaluation of neural network based approach for sentiment classification of online reviews. *Journal of King Saud University Computer and Information Sciences*, *28*(1), 2-12. https://doi.org/10.1016/j.jksuci.2014.03.024

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica, 10(5), 1122–1136.

Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2019). A survey of sentiment analysis in social media: L. Yue et al. *Knowledge and information systems*, *60*(2), 617-663. https://doi.org/10.1007/s10115-018-1236-4

Zhai, Y., Song, X., Chen, Y., & Lu, W. (2022). A study of mobile medical app user satisfaction incorporating theme analysis and review sentiment tendencies. *International journal of environmental research and public health*, *19*(12), 7466. https://doi.org/10.3390/ijerph19127466

Zhang, J., Li, D., Wang, X., Wu, Z., & Lyu, Q. (2025). Application and challenges of DeepSeek in primary care in China. *Frontiers in public health*, *13*, 1721308. http://dx.doi.org/10.3389/fpubh.2025.1721308

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, *4*(11), 218.

Corresponding author
Author Name can be contacted at: semmy@unklab.ac.id