Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Computer Science and Computational Intelligence 2025 (ICCSCI 2025)

Comparison of Traditional Machine Learning and Deep Learning Models for Tweet Sentiment Analysis

Jonathan Alvios[a], Hanzen Jonathan[a], Richardo Kaka Widjaja[a], Almuzhidul Mujhid[a], Hidayaturrahman[a]

[a]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia*

Abstract

Sentiment analysis on social media platforms has become a crucial task in understanding public opinion and emotional trends. This paper presents a comparative study between traditional machine learning models — Naive Bayes, Logistic Regression, and Random Forest — and deep learning models such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) for classifying sentiment in tweets. We use a publicly available Twitter sentiment dataset, applying consistent preprocessing, vectorization, and evaluation metrics across models. Experimental results indicate that while traditional models are faster to train and easier to interpret, deep learning models, particularly LSTM, demonstrate superior performance in capturing contextual sentiment. The findings offer insights into the trade-offs between model accuracy, training time, and complexity for sentiment analysis applications.

© 2025 The Authors. Published by ELSEVIER B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 10th International Conference on Computer Science and Computational Intelligence 2025

Keywords : Sentiment Analysis, Twitter, Naive Bayes, LSTM, CNN, Machine Learning, Deep Learning

1. Introduction

1.1 Background

We are now living in modern era, where social media has become a place for everyone for everyone to pour out their thoughts. Indirectly, social media become an important role for most people. One of the most famous social media in the world is Twitter (known as ‘X’ for now) which produces millions of tweets every day which contains user’s sentiment[1]. Sentiment Analysis on Twitter gained significant attention in recent years due to its potential in understanding public opinion, predicting, trends, and help for decision making[2][3][4]. Based on that, researchers conduct a survey related to sentiment analysis and it was proven that sentiment analysis is very important for

1877-0509 © 2025 The Authors. Published by ELSEVIER B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 10th International Conference on Computer Science and Computational Intelligence 2025

businesses, governments, and individuals with extracting sentiment from various online sources[4][5][6]. Various machine learning algorithms have been used for this task, including traditional models such as Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), and some deep learning approaches. Research related to deep learning models in analyzing sentiment is growing very fast from year to year. Several find that deep learning models outperform traditional models . Meanwhile there are also research that find traditional models, with proper feature engineering can achieve comparable results. Although deep learning models such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) shown superior performance in many studies, traditional models are often more efficient and easier to interpret. Several studies have tried to combine deep learning models into hybrid models that combine various architectures such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) or other models such as Deep Belief Networks (DBN) and Deep Neural Networks (DNN)[5][7][8]. The results shown are also positive, because each model has its own advantages in handling certain features. In other cases, there are also those who use sentiment lexicons for reviews of e-commerce products to strengthen emotional features in review texts[9]. However, there is still a research gap in understanding how the performance of these various models compared under different dataset conditions, such as large vs small data amounts, data with high noise, or data with different word representations (TF-IDF vs word embeddings)[1][3][10].

1.2 Research Questions

- How Naive Bayes, Logistic Regression, Random Forest, and some deep learning models compare in accuracy and efficiency in tweet sentiment analysis?
- Which model has the best trade-off between speed and accuracy with different variation of datasets in size and quality?

1.3 *Objectives*

This study aims to compare the performance of several traditional methods and deep learning models in tweet sentiment analysis. Specifically, the objective of this study are:

- Analyzing the performance of NaiveBayes, Logistic Regression, Random Forest, and several deep learning models in tweet sentiment classification[5][7][8].
- Evaluate the accuracy, F1-score and training time of each model to determine the most effective method[3][10][11].
- Identifying the advantages and disadvantages for each model tested in handling Twitter text data[4][12][13].
- Provides recommendations on the most optimal models for sentiment analysis on Twitter data, which can be used in various applications such as public opinion monitoring, brand analysis, and data-driven decision making[6][14].

2. Literature Review

2.1 Previous Research

A study by Syahputra, Yanris, and Irmayani compares two machine learning algorithms, Support Vector Machine (SVM) and Naïve Bayes for sentiment analysis on Twitter regarding the Peduli Lindungi application [15]. The dataset used consisted of 4,782 tweets. The researchers compared the accuracy and processing time of both algorithms. Using the k-fold test method, SVM algorithm achieved an accuracy of 86%, while the Naive Bayes achieved 85%. With 8020 data split method, SVM algorithm achieved 79% and Naive Bayes with 80%. The SVM algorithm required 84.06 seconds of processing time, whereas Naive Bayes was significantly faster with only 0.0094 seconds. Although both algorithms achieved the same level of accuracy, Naive Bayes required faster processing time.

A study conducted by Rodrigues, Fernandes, and others evaluated various classification algorithms and deep learning models for spam detection and sentiment analysis tasks [16]. For spam detection, the classification algorithms used include decision tree, logistic regression, multinomial Naive Bayes, support vector machine (SVM), random forest, and Bernoulli Naive Bayes. For sentiment analysis, the classification algorithms used include

stochastic gradient descent (SGD), SVM, logistic regression, random forest, and Naive Bayes. For both tasks, the deep learning models used were simple Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), bidirectional LSTM (BiLSTM), and 1D Convolutional Neural Network (CNN). The spam detection dataset consisted of 5,572 entries and the sentiment analysis dataset used consisted of 31,015 tweets. The best classification algorithm and deep learning models based on accuracy for each task are: for spam detection, the multinomial Naïve Bayes algorithm achieved an accuracy of 97.78%, while the LSTM model achieved 98.74%. For the sentiment analysis, the SVM achieved an accuracy of 70.56%, and the LSTM model with the accuracy of 73.81%.

2.2 Models

Sentiment analysis refers to the task of identifying emotions conveyed through text [4]. It involves analyzing large volumes of text to classify whether the content is positive, negative, or neutral. Sentiment analysis falls within the scope of natural language processing (NLP), which is a subfield of artificial intelligence. It does not merely involve recognizing specific words that carry positive or negative connotations, but rather understanding the overall tone of a statement to assess its sentiment [4]. There are lots of algorithms used for this specific task, from traditional machine learning algorithms to deep learning models.

Naive Bayes Classifier is one of the machine learning algorithms used for classification tasks. It is based on the application of Bayes Theorem. Naive Bayes has an explicit and strong theoretical foundation that ensures optimal induction under a set of explicit assumptions. The algorithm is fast and easy to implement, with a simple yet effective structure [13]. A key characteristic of this method is its strong (naïve) assumption of the independence between each feature. Study by Syahputra shows that Naive Bayes can achieve similar accuracy with way faster processing time compared to SVM algorithm [15].

Another widely used method is Logistic Regression. Logistic Regression is an algorithm used to obtain odds ratios through the presence of more than one independent variable [14]. It allows for the simultaneous analysis of multiple independent variables while minimizing the influence of confounding variables. Unlike linear regression, logistic regression produces a binomial outcome as the response variable. The classification is derived from sigmoid function [16].

Random Forest (RF) is another machine learning algorithm designed for making predictions. RF constructs a large number of decision trees during training and combines their output to make final predictions. RF is well-suited for medium to large datasets. In the study conducted by Schonlau, models using the RF algorithm achieved higher accuracy rates compared to both logistic and linear regression. RF is considered one of the best-performing machine learning algorithms due to its ability to adapt to nonlinearity within the data, enabling it to produce better predictions compared to linear regression [17].

Another commonly used algorithm is the Support Machine Vector (SVM). SVM is a supervised learning algorithm that analyzes data and recognizes its pattern. SVM is used for both classification and regression tasks. It works by finding the most optimal separator space between different classes within the dataset [15]. SVM is faster and performs better compared to modern approaches, such as neural networks, when it is trained with fewer data [16]. This makes SVM work well with a small to medium size dataset. SVM has been proven to outperform other traditional algorithms, although it is slower in processing speed compared to Naive Bayes, as shown in study by Syahputra [15].

Deep learning is a subset of machine learning based on artificial neural networks. In contrast to traditional machine learning approaches, where features are defined and extracted either manually or by making use of feature selection methods. Deep learning models automatically learn and extract features, thus improving accuracy and performance [10]. A key characteristic of deep learning is the presence of hidden layers during the data processing stage. Some experts argue that a network can be considered "deep" if it contains more than one hidden layer, whereas others believe it necessary for a model to have many hidden layers for it to be considered as one [18]. Two commonly used types of artificial neural networks are Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). CNN is applied in tasks such as image recognition, video analysis, and natural language

processing tasks. Whereas LSTM is used for tasks involving sequential data such as text classification and generation, speech recognition, music composition, and time series prediction [18].

A Deep Neural Network (DNN) is one of the deep learning models composed of neural networks structured in multiple layers that are designed to process data. These layers consist of interconnected neurons, and what distinguishes DNNs is their ability to extract complex patterns from data, making them well-suited for classification tasks [10]. Recurrent Neural Network (RNN) is another deep learning model tailored to handle sequential data, such as text, which contains temporal and sequential dependencies. RNN maintains a hidden state that retains information from the previous step in the sequence. This makes RNN particularly suitable for sentiment analysis, where the word order and context provided by preceding words greatly influences the meaning of a sentence [10].

3. Methodology

This study adopts a quantitative comparative experimental approach, aiming to evaluate and compare the performance of traditional machine learning models (SVC, Naive Bayes, Logistic Regression, and Random Forest) and deep learning models (such as LSTM and GRU) for sentiment classification analysis on Twitter data [3][5] [7]. The research focuses on measuring both the effectiveness (accuracy, precision, recall, F1-score and etc) and efficiency (training and inference time) of each model. The method process will be included accordingly with Fig.1.

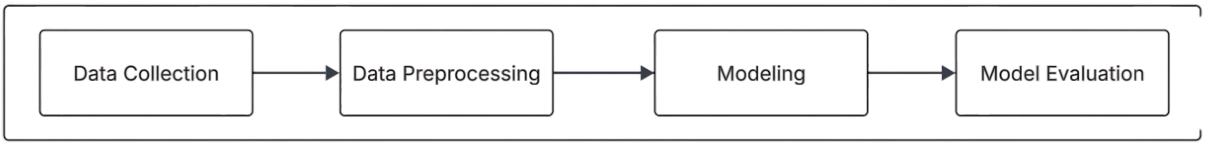

Fig. 1. Methodology Process

3.1 Data Collection

The study uses secondary data in the form of publicly available Twitter datasets that are already labeled with sentiment classes (positive, negative, neutral). The datasets considered include Sentiment140, Twitter US Airline Sentiment Dataset, Other relevant datasets from platforms such as Kaggle [19].

Criteria for dataset selection include a sufficient number of labeled tweets (at least 10,000 entries), balanced or near-balanced sentiment class distribution, publicly available for academic research[1][3].

The datasets will be obtained by downloading from publicly available repositories. Initial exploration and preprocessing will be carried out to ensure the data quality. Unwanted entries such as duplicates, irrelevant text, or incomplete labels will be removed.

3.2 Preprocessing Data

Before modeling, the data will undergo a preprocessing stage, which is crucial for improving text quality and eliminating noise that could affect model performance[4][5][7]. The preprocessing steps include converting all text to lowercase (lowercasing), removing noise such as URLs, hashtags, emojis, numbers, and special characters, and performing tokenization to split the text into individual words or tokens. Common words that do not contribute meaningful information to sentiment analysis, such as “is,” “the,” and “and,[4][5]” are removed in the stopword removal step. Additionally, stemming or lemmatization are applied to reduce words to their root form (e.g., “going” becomes “go”). Sentiment labels are also encoded into numerical format through label encoding. To ensure data integrity, text uniqueness analysis is conducted by identifying outliers or texts with extreme lengths, and duplicate or empty texts are removed [1]. The processed text is then transformed into numerical representation using techniques that vary based on the type of the model. Traditional models use Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF), whereas deep learning models utilize pre-trained word embeddings such as GloVe and Word2Vec[20].

3.3 Modeling

In this research, six machine learning and deep learning models were developed to perform sentiment classification on Twitter data[3][7][8]. The traditional models included Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVC), and Random Forest (RF). For the deep learning approach, we implemented Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) neural networks.

All models were trained on the same preprocessed dataset. For the traditional models, tweets were vectorized using TF-IDF to convert text into numerical features. Each classifier was then trained on these feature vectors. Naive Bayes served as a baseline probabilistic model, Logistic Regression was chosen for its efficiency on linear problems, SVC was applied due to its effectiveness on high-dimensional data, and Random Forest was used as an ensemble method that combines multiple decision trees[13][14][15][17].

The deep learning models (LSTM and GRU) were implemented using the Keras API. Tweets were tokenized and padded into sequences, then passed through an embedding layer followed by recurrent layers (either LSTM or GRU) and a dense output layer with a softmax activation for classification. The models were trained using categorical cross-entropy loss and the Adam optimizer, with dropout layers included to reduce overfitting[13][14][17].

All models were trained using an 80:20 split for training and testing datasets, and hyperparameters were tuned through experimentation.[7][10][17].

3.4 Model Evaluation

To evaluate model performance, we used standard classification metrics: Accuracy, F1-score, and training time. These metrics provide insight into how well each model classifies tweets into sentiment categories (positive, negative, and neutral). Table 1 & 2 presents a summary of the results.

Table 1. Model Performance Comparison on noisy data (rounding to four decimal places).

Model	Accuracy	F1-Score	Training Time (sec)
SVC	0,6896	0,6904	363
Naive Bayes	0,5919	0,5637	**1**
Logistic Regression	0,6759	0,6770	20
Random forest	0,6470	0,6417	363
LSTM	0,7076	0,7081	1504
GRU	**0,7098**	**0,7094**	1283

Table 2. Model Performance Comparison on cleaned data (rounding to four decimal places).

Model	Accuracy	F1-Score	Training Time (sec)
SVC	0,7006	0,7003	248
Naive Bayes	0,6036	0,5842	**1**
Logistic Regression	0,6861	0,6858	20
Random forest	0,6758	0,6742	570
LSTM	**0,7170**	**0,7164**	1860
GRU	0,7114	0,7096	1973

4. Result & Discussion

4.1 Model Performance on Noisy Data

Table 1 present the performance comparison of all models on noisy data. Noisy data in this case refers to condition of data with minimum data preprocessing. GRU model perform highest accuracy (0,7098) and F1-score (0,7094) followed closely by LTSM. Among traditional models, SVC perform highest accuracy (0,6896) and F1-score (0,6904). Naive Bayes has the worst performance in accuracy and F1-score of all models but has the fastest training time. As we can see on the table, the deep learning models perform better in accuracy and F1-score but has the longest training time, means had higher computational costs.

4.2 Model Performance on Cleaned Data

Table 2 shows the performance of same models on cleaned data. Cleaned data in this case refers to condition of data with maximum data preprocessing. Overall, all models improved in accuracy and F1-score, with LTSM achieve the highest accuracy (0,7170) and F1-score (0,7164) followed by GRU. With cleaned data, the deep learning models perform better in accuracy and F1-score, but the traditional models resulted a big improvements compared to Table 1. In terms of training time, the deep learning models has the longest training time.

4.3 Comparative Analysis

From the result, both of the table show that deep learning models out performing the traditional models in terms of accuracy and F1-score, but traditional models show more improvement than deep learning models. Deep learning models do not show significant improvement because deep learning model do not require a lot of data preprocessing. In terms of training time, traditional models show much faster than deep learning models, which mean deep learning model require more computational cost.

These finding show that traditional model has a fast training time so it is suitable for low computational resource environment. Deep learning models provide superior accuracy and F1-score so it is suitable for applications where accuracy is critical. Additionally, with good data preprocessing, the results of traditional models will be good and it will be closer to the accuracy of deep learning model.

5. Conclusion

This study presents a comparative analysis between traditional machine learning models (Naive Bayes, Logistic Regression, Support Vector Machine, Random Forest) and deep learning models (LSTM and GRU) for sentiment analysis on Twitter. Experiments conducted with two data conditions, noisy and cleaned to evaluate model accuracy, F1-score, and training time.

The result shows that deep learning models outperform traditional models in both accuracy and F1-score so it is suitable for applications where accuracy is critical. However, traditional models show more significant performance gains after data cleaning, which reflects to strong preprocessing techniques. Traditional models required less training time in both noisy and cleaned data, demonstrate its suitability for limited resources.

Overall, this study concludes that while deep learning models offer superior performance, traditional models remain practical for limited computational sources and lightweight applications. Future research can explore hybrid model approaches and upcoming deep learning model.

References

[1] Wahyuningsih, T., Chen, S.C., et al. (2024). *Analyzing sentiment trends and patterns in Bitcoin-related tweets using TF-IDF vectorization and k-means clustering.* EduLearn, 18(1), 173-184.

[2] Awangga, R.M. (2017). *A hybrid CNN-LSTM model with word-emoji embedding for improving Twitter sentiment analysis on Indonesia's PPKM policy.*

[3] Fiok, K., Karwowski, W., Gutierrez-Franco, E. (2022). *Comparison of model performance and explainability of predictions in sentiment analysis.* Energies, 15(21), 8079.

[4] Taherdoost, H., & Madanchian, M. (2023, February 1). Artificial Intelligence and Sentiment Analysis: A Review in Competitive Research. *Computers*. MDPI

[5] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1253.

[6] Ali, S., & Kabir, M. A. (2024). A review of sentiment analysis: tasks, applications, and deep learning approaches. *Journal of Ambient Intelligence and Humanized Computing*.

[7]. Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., & Rehman, A. (2017). Sentiment analysis using deep learning techniques: A review. *International Journal of Advanced Computer Science and Applications : IJACSA, 8.*

[8] Sahoo, C., Wankhade, M., & Singh, B. K. (2023). Sentiment analysis using deep learning techniques: a comprehensive review. *International Journal of Multimedia Information Retrieval*, 12(1), 41.

[9] Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE Access: Practical Innovations, Open Solutions, 8*, 23522–23530.

[10] Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. In *arXiv [cs.CL]*.

[11] Hussein, D. M. E.-D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences, 30*(4), 330–338.

[12] Kayed, M., Díaz-Redondo, R. P., & Mabrouk, A. (2023). Deep learning-based sentiment classification: A comparative survey. *arXiv*.

[13] Taheri, S., & Mammadov, M. (2013). Learning the naive bayes classifier with optimization models. *International Journal of Applied Mathematics and Computer Science*, 23(4), 787–795

[14] Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 24(1), 12–18.

[15] Syahputra, R., Yanris, G. J., & Irmayani, D. (2022). SVM and Naïve Bayes Algorithm Comparison for User Sentiment Analysis on Twitter. *Sinkron*, 7(2), 671–678.

[16] Rodrigues, A. P., Fernandes, R., Aakash, A., Abhishek, B., Shetty, A., Atul, K., … Shafi, R. M. (2022). Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques. *Computational Intelligence and Neuroscience*, *2022*.

[17] Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*, 20(1), 3–29.

[18] El-Hadi, M. (2022). Artificial Intelligence vs. Machine Learning vs. Deep learning. What is the Difference? مجلة الجمعية المصرية لنظم المعلومات وتكنولوجيا الحاسبات, 29(29), 82–84.

[19] Singh, U. (2023). A comparative study of sentiment analysis techniques: machine learning vs. deep learning. *International Journal for Scientific Research & Development*, 11(8), 80016.

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