Buletin Ilmiah Sarjana Teknik Elektro
Vol. 8, No. 3, June 2026, pp. 736-745
ISSN: 2685-9572, DOI: 10.12928/biste.v8i3.14373

A Novel Slang and Formal Text Classification with Data Exploration and Optimized Deep Learning Models

Hoger K. Omar
College of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq

ARTICLE INFORMATION

Article History:

Received 29 August 2025
Revised 25 October 2025
Accepted 04 June 2026

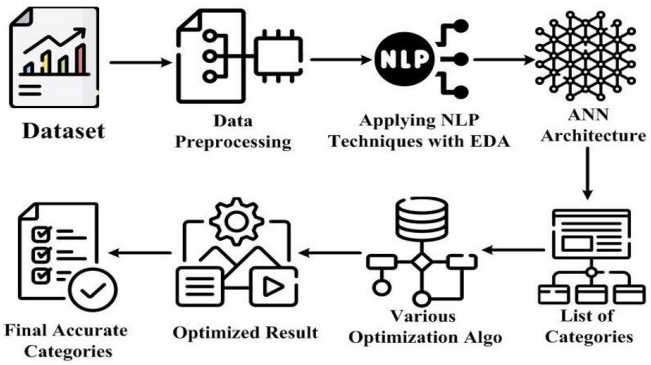
Keywords:

Exploratory Data Analysis;
Hyperparameter Algorithms;
Text Categorization;
Slang Classification;
Deep Learning

Corresponding Author:

Hoger K. Omar,
College of Computer Science and Information Technology,
University of Kirkuk, Kirkuk, Iraq.
Email:
hogeromar@uokirkuk.edu.iq

ABSTRACT



Automated text classification involves applying artificial intelligence algorithms to classify text documents into predefined categories. Hence, developing a high-accuracy text categorization model is a significant task, especially in unstructured narratives such as research papers, medical documents, and news articles. This study examines the application of an artificial neural network (ANN) algorithm for categorizing formal and slang English language with the capabilities of popular deep learning frameworks such as TensorFlow and Keras. First of all, the dataset's features were examined through exploratory data analysis (EDA) methods to enhance understanding. Furthermore, the study emphasizes the use of several preprocessing techniques to address the challenge presented by the informal writing style. In addition, adding a list of common English abbreviations greatly improved the accuracy and effectiveness of classifying text. Lastly, the work involves using multiple hyperparameter optimization approaches for further enhancement. The proposed techniques effectively mitigated the impact of heterogeneous and noisy data in both formal and informal language by achieving an improvement of approximately 10% in overall classification accuracy. Additionally, the study contributes to an advancement in the field of text mining and offers practical guidance for optimizing deep learning models in the domain of English text categorization.

Document Citation:

H. K. Omar, "A Novel Slang and Formal Text Classification with Data Exploration and Optimized Deep Learning Models," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 3, no. 3, pp. 736-745, 2026, DOI: 10.12928/biste.v8i3.14373.

1. INTRODUCTION

Indeed, most digital data is in the form of text and often unstructured or semi-structured. Consequently, to make data valuable for decision-making, the categorization of this textual data has become indispensable [1]. Recently, text mining and analysis have gained increasing attention due to the large volume of digital data generated from social media, blogs, and public libraries. [2]. Text classification can take on different forms, including binary categorization, where a text is assigned to one of two opposing groups. Multi-class categorization, where a text is assigned to one of several categories, or multi-label categorization, where a text can be assigned to zero, one, or multiple groups. Nowadays, multi-class techniques are widely used in both academic research and practical applications. Additionally, many multi-label approaches utilize problem transformation methods to convert a multi-label problem into multiple binary classification tasks [3]. Text categorization is essential in several domains such as information extraction, retrieval, text mining, and text analytics [4]. Figure 1 shows a general overview of the text categorization pipeline.

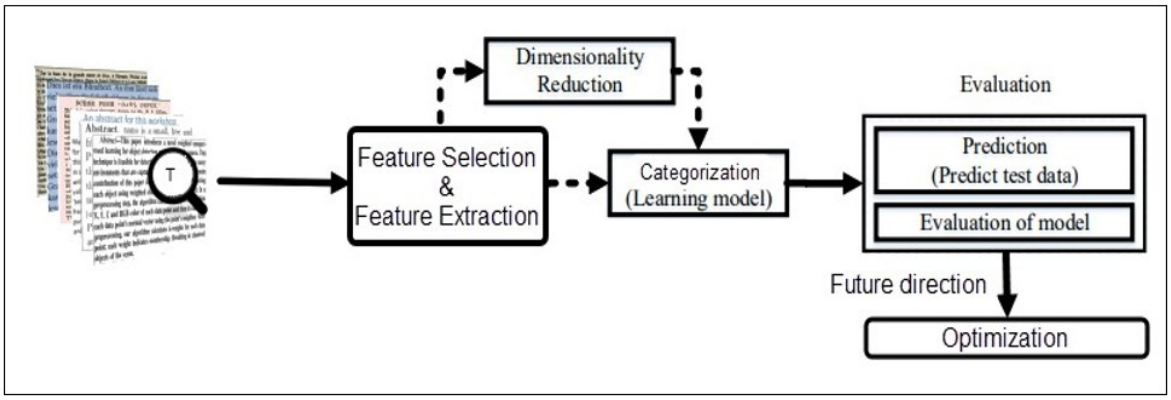


Figure 1. General overview of the text categorization pipeline

Over the years, the task of text classification has seen the development of different algorithms, broadly classified into two main groups: traditional machine learning (ML) and deep learning (DL) [5]. The DL refers to the latest technological advancement in the ML field. DL has become widespread in our everyday lives by providing solutions that were previously considered the domain of scientific inquiry [6]. The widespread utilization of deep learning has opened up opportunities for its application in various domains, including text categorization, machine translation, recommendation systems, and sentiment analysis [7]. The main reason for using deep learning algorithms is that they perform exceptionally well when dealing with highly complex data or when large amounts of training examples are available [8]. Thus, the current efforts in deep learning focus on improving how text is represented to boost the effectiveness of text categorization. [9]. Supervised deep learning algorithms have shown a tendency to make predictions with excessive confidence which is often described as catastrophic overconfidence [10]. Text categorization is a form of supervised learning that involves assigning natural language text items to predefined categories. Various classifiers often employ supervised ML and DL algorithms for this purpose.

The presence of noise features in the high-dimensional feature space reduces the accuracy of the classification. To address this, feature selection plays a crucial role in diminishing the dimensionality of the feature space and improving accuracy in text classification problems [11]. As well as the presence of noise in text data also reduces the accuracy of the classification because the accuracy of the extracted knowledge relies on the quality of the data employed [12]. However, the informal writing style employed by many users shows many challenges for various natural language processing (NLP) applications such as text categorization, chatbots, and sentiment analysis. This is because these applications are usually trained on clean or well-structured text [13]. Therefore, the preprocessing of noisy and standard text differs significantly because noisy text lacks standardized rules or patterns, unlike standard text, which adheres to language standard formats [14]. Hence, this work addressed the problem of heterogeneous, slang, and noisy textual data by utilizing various preprocessing pipeline [15]. Different techniques in text pre-processing and NLP can normalize unstructured/structured text data. NLP, as a subset of Artificial Intelligence (AI), empowers machines to comprehend spoken/written human language [16]. NLP utilizes diverse methods to interpret uncertainties in textual data. The utilization of NLP has revealed its essential role in effective data categorization [17]. A core objective in NLP is the conversion of text into numeric vectors [18]. But in this work, many other tasks have

been assigned to NLP, starting from preprocessing to employing techniques that enhance accuracy in collaboration with the Keras framework.

Noteworthy, this work utilized the Exploratory Data Analysis (EDA) which is a data analysis approach that frequently employs visual methods to succinctly summarize the principal characteristics of the data. The primary focus of EDA is to discern insights that go beyond formal models or hypothesis testing. It differs from initial data analysis (IDA), which is dedicated to validating assumptions vital for the model by assessing hypotheses, handling missing values, and adjusting variables as necessary. Notably, EDA includes the scope of IDA [19]. In a nutshell, the study proposed a text categorization approach in a Keras deep learning environment by combining the benefits of EDA, advanced preprocessing techniques, and various optimization algorithms to achieve optimal outcomes.

The goals of this study are outlined through the following research questions:

1. How can exploratory data analysis (EDA) techniques be applied to understand and refine textual data in order to improve the accuracy of text categorization models?
2. How can Natural Language Processing (NLP) techniques be effectively integrated with the Keras deep learning framework to enhance text categorization performance?
3. What preprocessing strategies are most effective in reducing the impact of noisy data on the performance of text categorization models?
4. How can text categorization models be improved to accurately process and classify texts written in informal English or containing slang expressions?

This work is organized as follows. Section 2 provides an overview of the related-work. Section 3 explained in detail the statistical method used in exploratory data analysis. Section 4 offers an in-depth exploration of our contributions. In Section 5, comprehensive information regarding the experiments and results is provided. Lastly, Section 6 outlines the conclusions.

2. LITERATURE REVIEW

In recent times, a substantial volume of scholarly publications has emerged concerning text categorization. These works have presented various methodologies for building efficient systems. This section will concentrate on testing the most esteemed contributions within this domain. E. I˙lgun *et al*, 2025 [20] Identified patterns in crime frequency across different time intervals, crime types, and police districts. The researchers have used diverse ML algorithms in order to classify and predict the type of crime, which include XGBoost, CatBoost, Random Forest, Decision Tree, MLP, KNN, and Logistic Regression. In addition, time series analysis was demonstrated to predict the future crime occurrences with both statistical models (Holt-Winters Exponential Smoothing and deep learning LSTM and BiLSTM) on a large number of police districts. G. G. Ro‘ziyeva *et al*, 2025 [21] examined the problem of social media text classification. The article was founded on short articles which were written in a conversational way. These posts also use slang, abbreviations, emojis and code switching and are therefore difficult to classify. Such type of data was experimented with machine learning, NLP and deep learning, in the research. It also proved that the context aware models and transfer learning may be used to improve the accuracy in the situation when the quantity of the labeled data is lower and that the language on the social sites is changing quickly. N. Hidayani *et al*., 2025 [22] studied how deep learning can be applied to sentiment analysis of informal language and slang social media texts. The experiment employed informally written comments and posts in Indonesian as the medium of testing the role of non-standard language in sentiment classification. A number of neural network models were used such as the RNN LSTM and CNN to determine the sentiment of each text. Preprocessing steps were also applied in the research to overcome the use of slang abbreviations and common spelling variants. The findings indicated that the LSTM model had the best accuracy. S. R. Naher *et al*, 2024 [23] explored the identification of Chittagonian slang in social media text to aid in the curbing of abusive language on the internet. The researchers provided a balanced dataset of 2,100 Chittagonian comments to improve the representation of this low resource language. Using both machine learning and deep learning models, i.e. LSTM and CNN models with Word2Vec embeddings, the research resulted in the highest accuracy of 76%. The results prove that deep learning methods can be used to advance the state of text classification in low-represented languages and enable useful content filtering on social networks. Z. Sun *et al*, 2024 [24] examined how well large language models understand and process slang. The study found that GPT-based models outperform BERT-like models in several areas: recognizing slang in everyday sentences, identifying where and when it is used, and predicting slang meaning more accurately than literal word alternatives. The results suggest that GPT does not treat slang as a separate type of language but learns it as uncommon word meanings within context. The study also showed that all models struggle more with UK and modern slang due to limited data. M. Orosoo *et al*, 2023 utilized [25]

Presented the integration of qualitative techniques to classify and evaluate extensive amounts of information in English text. The Lion Optimization

Algorithm (LOA) is used to complete the fundamental components of high-quality English compositions. The classification of the acquired texts is carried out using a Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM). H. Maragheh *et al*, 2022 [26] Proposed a novel system known as Spotted Hyena Optimizer (SHO)-Long Short-Term Memory (SHO-LSTM) for Multi-Label Text Classification (MLTC) using the LSTM network and SHO algorithm. In the LSTM network designed for MLTC, the Skip gram scheme is applied to insert words into the vector space. Their new model utilizes the SHO algorithm to enhance the initial weight of the LSTM network. The enhancement in the proposed model in comparison to the traditional LSTM model scored better results across four different datasets. A. Zhang *et al*, 2022 [27] Suggested a text classification deep active learning system utilizing the BERT algorithm. They developed an instance selection strategy relying on Margin, Intra-correlation, and Inter-correlation (MII) derived from posterior probabilities. Utilized instances exhibit characteristics such as a small margin, low intra-cohesion, and high inter-cohesion. Results indicate superior accuracy compared to baseline methods. A. Kurniasih *et al*., 2023 [28] investigated whether text preprocessing is essential when using transfer learning for text classification. The findings showed that processes such as correcting slang, fixing spelling mistakes and case folding are not always required if the right word embedding technique and deep learning model are selected. With BERT embeddings applied to LSTM, BiLSTM and CNN models, the accuracy difference between cleaned and unprocessed text was very small. The best performance was reached using a combination of BERT embeddings and a CNN model. The authors suggested that future efforts should focus more on model and embedding selection rather than excessive preprocessing, though GRU and MLP models showed different behavior. S. Piscitelli *et al*, 2021 [29] Proposed a multilingual application that will sort tweets according to the information content. They used multilingual word embeddings to create a deep learning classifier to perform real-time classification of various languages. The method enables the system to be trained on one language and translates flawlessly on any other language using zero-shot inference with a reasonable performance. Rianto *et al*, 2021 [30] presented a new stemming method to overcome the preprocessing problem with non-formal Indonesian text. The purpose of the research was to make the process of text classification more accurate through the betterment of the stemming procedure. A Support Vector machine classifier was constructed and tested on a sample of 550 Indonesian text samples with both the suggested stemming algorithm and a conventional algorithm. The findings have indicated that the enhanced stemming method was more accurate in the way it scored 0.85 whereas the standard method scored 0.73. These results suggest that the presented stemming technique decreases classification errors and generates a more valid model to predict the text categories. L. Qing *et al*, 2019 [31] Presented a new neural network model of medical text classification. In the sentence representation, a convolutional layer is used to extract the features, and a bidirectional gated recurrent unit (BIGRU) is used to obtain the previous and successive features of the sentences. The fact that an attention mechanism is included helps to derive the sentence representation whereby weight is given to important words. In the case of document, the process takes document-representation of sentences obtained via sentence representation and then encodes the sentences back via the attention mechanism to produce the document representation with vital sentence weights. J. Wang *et al*, 2019 [32] Suggested a new deep learning approach to text document classification, which used a scope-based CNN. As opposed to the traditional CNN which is window based, their method, which is called Large-Scale Scope-Based CNN, does not rely on adjacent words to build a local feature. LSS-CNN uses scope convolution, combination optimization, and max pooling operations to elicit deeper local text data insights. The suggested model obtains the most useful local information on text documents in a systematized way. The paper also goes in detail to discuss good methods of computing scope-based information and parallel training methods to suit large-scale datasets. X. Wang *et al*, 2018 [33] Constructed developed collateral Convolutional Neural Networks (CNNs) which support disorder and use variable-length pooling. Besides, in creating the bidirectional Long Short-Term Memory (LSTM), they removed input/output gates. The given method was evaluated on four topic and sentiment classification benchmark data. Combining LSTM regional embeddings and convolutional data processing led to the best results. It is important to note that the method beats all preceding methods, including deep learning methods, when it comes to topic classification and sentiment classification.

Recent work in text classification has been more and more concerned with informal and noisy text, particularly as social media content has emerged. The previous research work mainly used the conventional machine learning algorithms with simple preprocessing methods that proved efficient in cases of formal language but exhibited poor results when working with slang, short form, misspellings and irregular grammar. These methods tended to be weak in data exploration and therefore resulted in models which were vulnerable to heterogeneous and unstructured text. The deep learning models that enhanced classification performance

with learning contextual relationships are CNNs, LSTMs and transformers. Nonetheless, most of these works still took clean input text, and it did not extensively deal with linguistic differences between formal and informal writing. Very few studies analyzed the contribution of exploratory data analysis or systematic abbreviation normalization as the element of the learning process.

Additionally, most prior models relied on default hyperparameters or basic tuning, which restricted their accuracy and generalization capability. This study builds on and extends existing work by combining three key contributions rarely integrated together:

1. Exploratory Data Analysis to understand data distribution and detect heterogeneity in formal and slang English.
2. Targeted preprocessing including abbreviation mapping and slang normalization to minimize noise.
3. Advanced hyperparameter optimization using RandomSampler, CMA-ES, Optuna and TPE to maximize deep learning performance.

Thus, in contrast to the past research that aimed to determine one of the improvements (e.g. just preprocessing or just model tuning), the state of the art of this paper proves that progressive incorporation of EDA, normalization and optimization produces regular accuracy. This makes the method a more holistic and efficient solution to categorize formal and informal written English, as well as recommends future applications to multilingual scenarios and real-world applications.

3. EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is a method that is applied to explore data, identify valuable and actionable insights, differentiate the relationship between explanatory variables, identify the error, and preliminarily select the appropriate models. It is a method of then representing the data with the help of descriptive statistics and graphical tools to create an all-encompassing picture of the data [34]. Moreover, EDA effectively summarizes key data point of datasets, including the number of rows and columns, missing data identification, data types, and preview. The process also involves cleaning up corrupted data through solving problems such as missing data, invalid type of data and wrong values [35].

The visualization process involves the representation of data distributions by bar charts, histograms, and box plots. Moreover, it involves the calculation and visualization of the correlation or relationships between variables, usually represented in a heatmap. In addition, it summarizes the statistical characteristics of the information, especially focusing on four basic points, namely measures of central propensity (including mean, mode, and median), measures of dispersion (standard deviation and variance), the form of the distribution, and recognition of outliers [36].

4. PROPOSED SYSTEM CONSTRUCTION

The proposed system architecture for document categorization begins with the collection of raw text data. This includes both professional and casual (slang) English documents from a range of sources including news, social media and forums. This helps the system generalize across different language styles and variations. After filtering the data, the system performs a comprehensive data preprocessing step where multiple tasks are executed. This involves cleaning (removing HTLM, emojis, special characters and punctuation), normalizing (standardizing casing and spellings), and tokenising the text (breaking text into tokens). Then, the tokens are stemmed or lemmatized to convert them to their base forms. Finally, stop words (words typically removed as they aren't significant to the data meaning) and unrelated characters are eliminated to clean the data. For removal of stop words, the paper used the NLTK access to the English stop words corpus and added other domain-specific stop words to enhance linguistic features in the corpus according to exploratory data analysis. Next, we apply a bespoke abbreviation list to get additional text for abbreviations and contractions (Characteristic acronym). For example, "idk" is expanded to "I don't know". This greatly assists the model's understanding of user-generated text. In particular, the abbreviation list is a collection of about 550 typical English abbreviations. The list was generated from open linguistic data and subsequently reviewed manually to ensure it was appropriate within the context.

After the preprocessing of the text EDA is conducted to analyze the class distributions, detect class imbalances, and visualize terms frequencies. EDA not only directs the feature engineering, but it also reports to such strategies as class weighting or data augmentation when necessary. The second step entails the use of NLP features extraction. The conventional statistical methods are used at the start of the feature extraction process. Therefore, TF-IDF is employed to determine the importance of words by comparing the frequency of words in a particular document to that of words in the entire dataset. Although TF-IDF is effective at pointing out critical words, it does not have the ability to identify relationships between the semantics or context. To address these drawbacks, the system relies on the GloVe pre-trained word embeddings to encode words in the

continuous vector space in dense vectors, and aid the model learn to recognize word similarities according to the usage patterns. Furthermore, for an even more sophisticated representation of language, contextualized embeddings from models such as BERT are employed. Contrary to the static embeddings, BERT retrieves the intended meaning of a word based on the words surrounding the word in a sentence. The system then moves to the design of a powerful Artificial Neural Network. This constitutes input layers to take tokenized or embedded vectors, embedding layers (when non-BERT embeddings are being used) and multiple hidden layers with ReLU activation functions to acquire more complex patterns. The dropout rates are added to prevent overfitting by disabling part of the neurons randomly. The output layer is a softmax classifier in the case of multi-class classification with the number of nodes equal to that of document categories (e.g., politics, business, sports, etc.). The ANN is trained using the categorical cross-entropy loss function and the Adam optimizer which is selected because of its adaptive learning features. The system employs superior hyperparameter optimization methods to enhance the performance of the models. In particular, hyperparameter spaces are explored with the help of Optuna, RandomSampler, CMA-ES, and TPE. Learning rate, batch size, the number of layers, neurons per layer, and dropout rates are optimized. This move will guarantee an architecture is neither underfitted nor overfitted and optimal generalization is achieved. Lastly, a significant number of key measures are used to measure the model's performance. Figure 2 describes the process of the proposed system architecture from data collection to final classification.

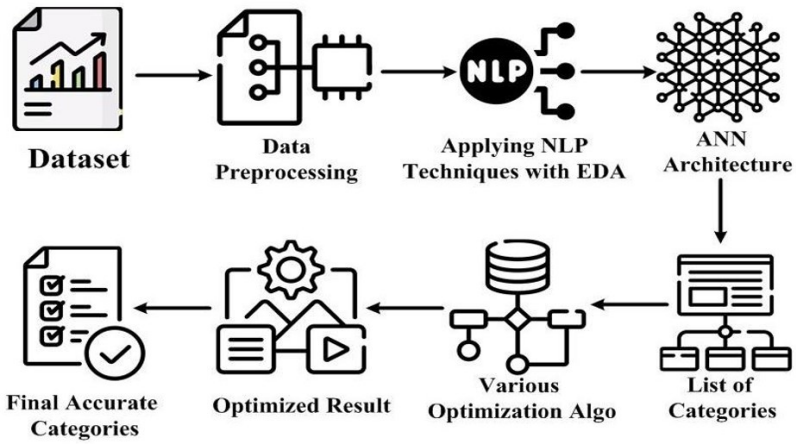


Figure 2. Proposed System Architecture

5. EXPERIMENTAL RESULTS

This section will provide a close summary of the experimental results that were obtained by the offered system of documents classification. The formal and informal English texts were used as a dataset to conduct the experiments. Accuracy of the system was used as a measure of system performance. In addition, the heatmap visualization was also used to analyze misclassifications. The findings indicated that the performance was evidently improved with every enhancement phase as depicted in the results subsection.

5.1. Data Collection

In this experiment, a real data was gathered in the form of different well-known English websites and sorted into five major categories: Business, Entertainment, Politics, Sports, and Technology. It contains approximately 2500 articles, which is a size of 5MB. During the process of data collection, the focus was on ensuring that there was diversity of text i.e. variety of content in various websites to fully test the potential of the model. The article in each record is long between 500 to 1000 words. That means that the designed model is quite universal and can process different types of text, such as articles, comments, reviews, and other contents. The use of such a comprehensive dataset improves the rigor and versatility of the model across the various text genres.

5.2. Results

This part shows the development of the model performance at various levels of improvement. It is concerned with the contribution that each technique made to enhancing the accuracy of text in English via a deep ANN model. Initially, the deep artificial neural network (ANN) model achieved an accuracy of 89.3% using standard preprocessing and text mining techniques such as data cleaning, tokenization, stop-word

removal, lemmatization, etc. By applying EDA, the system performance improved to 91.7%. EDA helped to discover patterns, imbalances, and outliers in the dataset and allowed more informed preprocessing and better feature representation. It provided a deeper understanding of the dataset structure and characteristics. In the EDA tests of this research, the graphical techniques are used to visually and diagrammatically summarize the data. Various graphical methods were applied using some Python libraries such as NumPy, plotly, pandas, and others. Figure 3 exhibits a heatmap showing the distributions of text clusters. Then, the useful addition of a standard English abbreviation dictionary and a slang list enabled us to overcome the difficulties associated with colloquial English found in the data. This has increased the level of accuracy to 94.8%. This finding demonstrates the necessity to consider superior language tricks, including the list of abbreviations to improve the performance of natural language processing systems. The words selected from a list are shown in Figure 4.

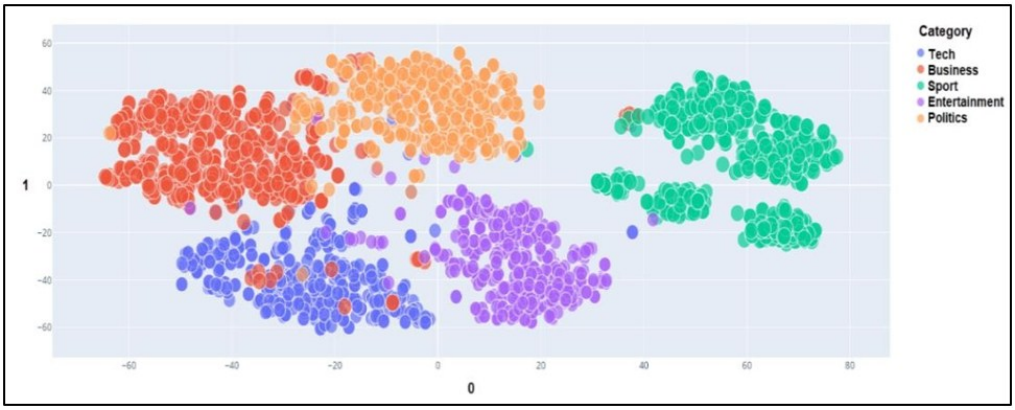


Figure 3. Heatmap illustrating the distributions of text clusters

```
abbreviations = {
    "$" : " dollar ",
    "€" : " euro ",
    "4ao" : "for adults only",
    "a.m" : "before midday",
    "a3" : "anytime anywhere anyplace",
    "aamof" : "as a matter of fact",
    "acct" : "account",
    "adih" : "another day in hell",
    "afaic" : "as far as i am concerned",
    "afaict" : "as far as i can tell",
    "afaik" : "as far as i know",
    "afair" : "as far as i remember",
    "afk" : "away from keyboard",
    "app" : "application",
    "approx" : "approximately",
    "apps" : "applications",
    "asap" : "as soon as possible",
    "asl" : "age, sex, location",
    "atk" : "at the keyboard",
    "ave." : "avenue",
    "aymm" : "are you my mother",
    "ayor" : "at your own risk",
    "b&b" : "bed and breakfast",
    "b+b" : "bed and breakfast",
    "b.c" : "before christ",
    "b2b" : "business to business",
    "b2c" : "business to customer",
    "b4" : "before",
    "b4n" : "bye for now",
    "b@u" : "back at you",
    "bae" : "before anyone else",
    "bak" : "back at keyboard",
    "bbbg" : "bye bye be good",
```

Figure 4. Exhibits a snapshot of some utilized abbreviations

For additional enhancement, several hyperparameter tuning strategies were used to additional optimize the deep ANN model. Fine-tuning plays a significant role in model performance, where parameters greatly affect the learning process. Techniques such as RandomSampler, Covariance Matrix Adaptation Evolution Strategy (CMA-ES), Optuna, and Tree-structured Parzen Estimator (TPE) were applied to identify the most effective configurations for the model. The outcomes of this step displayed that the optimized deep ANN model with RandomSampler achieved 96.0% accuracy, while CMA-ES and Optuna produced slightly higher

accuracies of 96.5% and 96.7% respectively. The greatest execution was achieved using TPE optimization with a final accuracy of 97.1%. Table 1 and Figure 5 show the rate of improvement in the results after applying each method. On the other hand, Figure 6(a) shows the training accuracy and loss curves, and Figure 6(b) shows the validation accuracy and loss curves.

Table 1. The results of the ANN algorithm before and after using some hyperparameter optimization.

No.	Method	Type of techniques	Acc. results
1	Deep ANN	Before any Techniques	89.3
2	Deep ANN	With EDA	91.7
3	Deep ANN	With EDA and Abbreviation List	94.8
4	Optimized Deep ANN	With RandomSampler	96.0
5	Optimized Deep ANN	With CMA-ES	96.5
6	Optimized Deep ANN	With Optuna	96.7
7	Optimized Deep ANN	With TPE	97.1

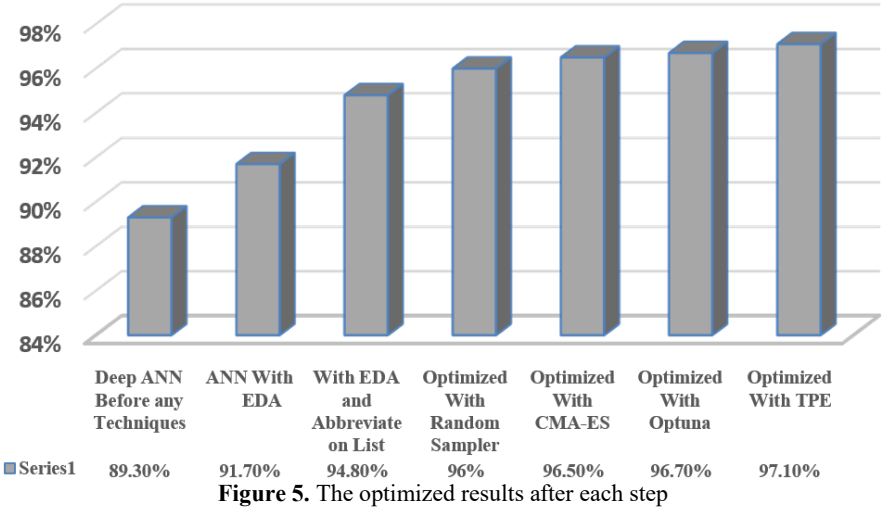
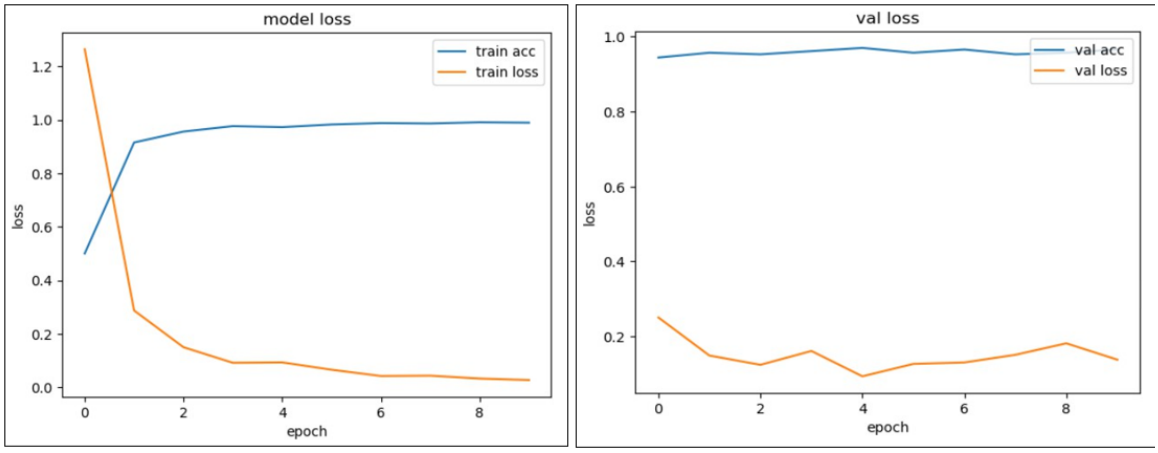


Figure 5. The optimized results after each step



(a) (b)

Figure 6. (a) raining Acc. and Loss Curve (b) Validation Acc. and Loss Curve

6. CONCLUSION

This study showed the effectiveness of applying deep learning with advanced data exploration and preprocessing techniques for classifying both formal and slang English texts. The system showed significant improvement from tackling issues such as informal language and diverse data through exploratory data analysis, text normalization, and abbreviation management. Also, the use of different techniques for hyperparameter tuning led to improved accuracy of the deep learning models, highlighting the need for tuning in deep learning workflows. The results have proven that each further development from EDA to abbreviation processing to hyperparameter tuning resulted in better classification scores. In this paper, the use of

sophisticated samplers such as Random Sampler, CMA-ES, Optuna, and TPE again demonstrated the effectiveness of hyperparameter tuning while working with deep learning classification models. The final accuracy has been tuned from 89.3% to 97.1% ensuring the success of implemented techniques which can efficiently deal with both formal as well as informal English text. Future work could test the approach with multilingual data. As well as testing it in practice for news filtering, content moderation, or categorization of academic papers.

REFERENCES

[1] V. Dogra, S. Verma, Kavita, P. Chatterjee, J. Shafi, J. Choi, and M. F. Ijaz, "A Complete Process of Text Classification System Using State-of-the-Art NLP Models," *Computational Intelligence and Neuroscience,* pp. 1-26, 2022, https://doi.org/10.1155/2022/1883698.

[2] A. Dhar, H. Mukherjee, N. S. Dash, and K. Roy, "Text categorization: past and present," *Artificial Intelligence Review*, vol. 54, no. 4, pp. 3007-3054, 2021, https://doi.org/10.1007/s10462-020-09919-1.

[3] R. A. Sinoara, J. Camacho-Collados, R. G. Rossi, R. Navigli, and S. O. Rezende, "Knowledge-enhanced document embeddings for text classification," *Knowledge-Based Systems*, vol. 163, pp. 955-971, 2019, https://doi.org/10.1016/j.knosys.2018.10.026.

[4] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artificial Intelligence Review,* vol. 52, p. 273–292, 2019, https://doi.org/10.1007/s10462-018-09677-1.

[5] J. T. Pintas, L. A. Fernandes, and A. C. B. Garcia, "Feature selection methods for text classification: a systematic literature review," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 6149-6200, 2021, https://doi.org/10.1007/s10462-021-09970-6.

[6] H. K. Omar, M. Frikha, and A. K. Jumaa, "Improving big data recommendation system performance using NLP techniques with multi attributes," *Informatica*, vol. 48, no. 5, 2024, https://doi.org/10.31449/inf.v48i5.5255.

[7] S. Guo, X. Li, and Z. Mu, "Adversarial machine learning on social network: A survey," *Frontiers in Physics*, vol. 9, p. 766540, 2021, https://doi.org/10.3389/fphy.2021.766540.

[8] H. K. Omar, M. Frikha, and A. K. Jumaa, "big data cloud-based recommendation system using NLP techniques with machine and deep learning," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 21, no. 5, pp. 1076-1083, 2023, https://doi.org/10.12928/telkomnika.v21i5.24889.

[9] H. Wu, S. Qin, R. Nie, J. Cao and S. Gorbachev, "Effective Collaborative Representation Learning for Multilabel Text Categorization," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5200-5214, 2022, https://doi.org/10.1109/TNNLS.2021.3069647.

[10] J. Van Landeghem, M. Blaschko, B. Anckaert and M. -F. Moens, "Benchmarking Scalable Predictive Uncertainty in Text Classification," in *IEEE Access*, vol. 10, pp. 43703-43737, 2022, https://doi.org/10.1109/ACCESS.2022.3168734.

[11] K. Thirumoorthy and K. Muneeswaran "Feature selection using hybrid poor and rich optimization algorithm for text classification," *Pattern Recognition Letters,* vol. 147, pp. 63-70, 2021, https://doi.org/10.1016/j.patrec.2021.03.034.

[12] H. Benhar, A. Idri, and J. L. Fernández-Alemán, "Data preprocessing for heart disease classification: A systematic literature review," *Computer Methods and Programs in Biomedicine,* vol. 195, p. 105635, 2020, https://doi.org/10.1016/j.cmpb.2020.105635.

[13] T. Baldwin and Y. Li, "An in-depth analysis of the effect of text normalization in social media," In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 420-429, 2015, https://doi.org/10.3115/v1/N15-1045.

[14] M. F. R. Abu Bakar, N. Idris, L. Shuib and N. Khamis, "Sentiment Analysis of Noisy Malay Text: State of Art, Challenges and Future Work," in *IEEE Access*, vol. 8, pp. 24687-24696, 2020, https://doi.org/10.1109/ACCESS.2020.2968955.

[15] D. Baktibayev, A. Serek, B. Berlikozha, and B. Rustauletov, "Resource-Efficient Sentiment Classification of App Reviews Using a CNN-BiLSTM Hybrid Model," *Buletin Ilmiah Sarjana Teknik Elektro,* vol. 7, no. 3, pp. 427-433, 2025, https://doi.org/10.12928/biste.v7i3.13954.

[16] P. Vashisth and K. Meehan, "Gender Classification using Twitter Text Data," *2020 31st Irish Signals and Systems Conference (ISSC)*, pp. 1-6, 2020, https://doi.org/10.1109/ISSC49989.2020.9180161.

[17] H. K. Omar, M. Frikha, and A. K. Jumaa, "PyTorch and TensorFlow Performance Evaluation in Big Data Recommendation System," *Ingénierie des Systèmes d'Information,* vol. 29, no. 4, pp. 1357-1364, 2024, https://doi.org/10.18280/isi.290411.

[18] R. Vinayakumar, K. P. Soman and P. Poornachandran, "Deep encrypted text categorization," *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 364-370, 2017, https://doi.org/10.1109/ICACCI.2017.8125868.

[19] E. S. Alamoudi and S. A. Azwari, "Exploratory Data Analysis and Data Mining on Yelp Restaurant Review," *2021 National Computing Colleges Conference (NCCC)*, pp. 1-6, 2021, https://doi.org/10.1109/NCCC49330.2021.9428850.

[20] E. G. İlgün and M. Dener, "Exploratory data analysis, time series analysis, crime type prediction, and trend forecasting in crime data using machine learning, deep learning, and statistical methods," *Neural Computing and Applications*, vol. 37, no. 18, pp. 11773-11798, 2025, https://doi.org/10.1007/s00521-025-11094-9.

[21] G. G. Ro 'ziyeva, B. I. Otaxonova, and M. E. Shaazizova, "Text Classification for Social Networks: Solving Short Text and Informal Language Problems," In *Conference on Internet of Things and Smart Spaces*, pp. 121-127, 2024, https://doi.org/10.1007/978-3-031-95296-8_11.
[22] N. Hidayani, T. Mantoro and M. A. Ayu, "Deep Learning Model for Sentiment Analysis in the Use of Informal Language and Slang On Social Media," *2024 10th International Conference on Computing, Engineering and Design (ICCED)*, pp. 1-5, 2024, https://doi.org/10.1109/ICCED64257.2024.10983073.
[23] S. R. Naher, S. Sultana, T. Mahmud, M. T. Aziz, M. S. Hossain and K. Andersson, "Exploring Deep Learning for Chittagonian Slang Detection in Social Media Texts," *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET*, pp. 1-6, 2024, https://doi.org/10.1109/ICECET61485.2024.10698491.
[24] Z. Sun, Q. Hu, R. Gupta, R. Zemel, and Y. Xu, "Toward informal language processing: Knowledge of slang in large language models," In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1683-1701, 2024, https://doi.org/10.18653/v1/2024.naacl-long.94.
[25] M. Orosoo, S. Govindasamy, N. Bayarsaikhan, Y. Rajkumari, G. Fatma, R. Manikandan, and B. K. Bala, "Performance analysis of a novel hybrid deep learning approach in classification of quality-related English text," *Measurement: Sensors*, vol. 28, p. 100852, 2023, https://doi.org/10.1016/j.measen.2023.100852.
[26] H. Khataei Maragheh, F. S. Gharehchopogh, K. Majidzadeh, and A. B. Sangar, "A new hybrid based on long short-term memory network with spotted hyena optimization algorithm for multi-label text classification," *Mathematics*, vol. 10, no. 3, p. 488, 2022, https://doi.org/10.3390/math10030488.
[27] A. Zhang, B. Li, W. Wang, S. Wan, and W. Chen, "MII: A Novel Text Classification Model Combining Deep Active Learning with BERT," *Computers, Materials & Continua*, vol. 63, no. 3, 2020, https://doi.org/10.32604/cmc.2020.09962.
[28] Manik, L. P. (2022). On the role of text preprocessing in BERT embedding-based DNNs for classifying informal texts. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2022, https://doi.org/10.14569/IJACSA.2022.01306109.
[29] S. Piscitelli, E. Arnaudo and C. Rossi, "Multilingual Text Classification from Twitter during Emergencies," *2021 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, USA, 2021, pp. 1-6, 2021, https://doi.org/10.1109/ICCE50685.2021.9427581.
[30] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation," *Journal of Big Data*, vol. 8, no. 1, p. 26, 2021, https://doi.org/10.1186/s40537-021-00413-1.
[31] L. Qing, W. Linhong, and D. Xuehai, "A novel neural network-based method for medical text classification," *Future Internet*, vol. 11, no. 12, p. 255, 2019, https://doi.org/10.3390/fi11120255.
[32] J. Wang, Y. Li, J. Shan, J. Bao, C. Zong and L. Zhao, "Large-Scale Text Classification Using Scope-Based Convolutional Neural Network: A Deep Learning Approach," in *IEEE Access*, vol. 7, pp. 171548-171558, 2019, https://doi.org/10.1109/ACCESS.2019.2955924.
[33] X. Wang and H. C. Kim, "Text Categorization with Improved Deep Learning Methods," *Journal of Information & Communication Convergence Engineering*, vol. 16, no. 2, 2018, https://doi.org/10.6109/jicce.2018.16.2.106.
[34] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian and A. Alothaim, "Exploratory Data Analysis and Classification of a New Arabic Online Extremism Dataset," in *IEEE Access*, vol. 9, pp. 161613-161626, 2021, https://doi.org/10.1109/ACCESS.2021.3132651.
[35] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, "Exploratory data analysis using Python," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 4727-4735, 2019, https://doi.org/10.35940/ijitee.L3591.1081219.
[36] A. Kulkarni and A. Shivananda. *Natural language processing recipes*. Apress. 2019. https://doi.org/10.1007/978-1-4842-4267-4.

AUTHOR BIOGRAPHY

Hoger K. Omar is currently an instructor at the University of Kirkuk and head of the Accreditation Section in the Quality Assurance Department / Presidency of Kirkuk University. His research interests include Big data, data mining, text classification, machine learning, artificial intelligence, distributed systems with Hadoop, recommendation systems, and natural language processing. He received a Ph.D. in Artificial Intelligence from the University of Sfax, Tunisia, in 2024.
Email: hogeromar@uokirkuk.edu.iq
ORCID: https://orcid.org/0000-0002-1942-5928
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57720377400