SLU JST
Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023], pp. 139-159 https://doi.org/10.56471/

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

# Effect of Hyperparameter Optimization on Deep Learning Models for Sentiment Analysis

Amos Orenyi Bajeh, Tijesuni Juliet Oladimeji, Ikeola Suhurat Olatinwo, Ghaniyyat Bolanle Balogun
Department of Computer Science, University of Ilorin, Ilorin, Nigeria
bajehamos@unilorin.edu.ng, oladimejitijesuni@gmail,com, olatinwo.is@unilorin.edu.ng, balogun.gb@unilorin.edu.ng

## Abstract

***Background***: *Huge textual and multimedia data are being generated daily on social media, online news publishing, and e-commerce platforms. Data mining has shown that valuable knowledge is hidden in large data, and this knowledge can be discovered using machine learning models and other data mining techniques. The modeling of machine learning algorithms for the mining of opinion from data relies on the goodness of the data and the model's set of parameters with which it learns the patterns in the data for future classification or prediction. This study investigated the impact of hyperparameter tuning on deep learning models for sentiment analysis.* ***Method****: Three prominent deep learning models: a convoluted neural network, long short-term memory, multilayer perceptron, and a hybridized convoluted neural network and long short-term memory, were considered. A random search technique, randomsearchcv, was used to optimize the hyperparameters of these models. Three prominent datasets: tweets on US airlines, the stock market dataset, and the IMDB movie review dataset. Two sets of the models were implemented: a set of optimized models in which their hyperparameters were tuned for optimal performance, and the other set of unoptimized models in which their hyperparameters were not tuned.* ***Results****:The results of these two sets were compared. Also, the result of the optimized models were compared to that of another study which used the same type of models and the IMDB movie review dataset. The results showed hyperparameter optimization marginally improved the performances of the models. Also observed is that the hybridized CNN-LSTM model outperformed the other single models.*

***Keywords****: sentiment analysis, deep learning, hyperparameters tuning.*

## 1. Introduction

The evolution of the Internet has effectively reduced the world to a global village by facilitating rapid communication between entities irrespective of geographical location through media such as blogs and various social media platforms. The Internet has also radically changed the way and manner commerce is conducted. Internet availability has liberalized communication such that users can express their thoughts and opinions about products, incidences, and any topic of interest.

The practice of extracting a writer's thoughts, sentiments, or attitude from written text is known as sentiment analysis or opinion mining. It employs natural language processing (NLP), data mining, and machine learning concepts to determine writers' opinions about products or topic of interest. This involves constructing a system to collect and monitor opinions about a brand, product, or service expressed in blog posts, comments reviews, forum discussions or tweets, and determining the polarity of the opinions whether they are negative, positive, or neutral. Social media monitoring, market research, brand monitoring, and customer service all rely heavily on sentiment analysis. E-commerce sites use online opinions to sell their products, and web customers read the product reviews before purchasing it. Businesses use sentiment analysis to analyse client feedback on their products and services on the internet (Padhye et al., 2018).

SLU JST
Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023],
pp. 139-159 https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

There are three approaches to extracting sentiment from text: lexicon-driven, machine learning-based, and a mix of the two (Ahmad et al., 2017). The lexicon-based technique generalizes the polarity of any given text using its dictionary of weighted words. Before conducting a classification task, the machine learning (ML) approach requires training an algorithm with pre-labeled data (training dataset) and then testing the algorithm with a test dataset which was not used to train the algorithm. The hybrid approach incorporates the best of both worlds: machine learning and lexicon-based approaches.

Deep learning (DL) is a branch of machine learning in which the algorithm is based on the structure of human brain, mimicking the physiology of the neural network in the human brain. Artificial Neural Network (ANN) is the foundation of deep learning. A neural network takes in data, recognizes patterns in the data, and then use the pattern learned from the data to predict the output for new input data not used in the initial dataset. Layers of neurons make up neural networks, and these neurons form the network's main processing unit. The input layer gets the data, the output layer predicts the outcome, and the hidden layer(s) learn the pattern embedded in the data. Deep learning is achieved by combining numerous layers in a neural network. DL models include convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term memory networks (LSTM), deep brief networks (DBN), Generative adversarial networks (GAN), and others. The performance of DL models significantly depends on how well they are tuned to learn the patterns in data. The tuning involves adjusting the values of some parameters referred to as hyperparameters in the models. These hyperparameters include learning rate, activation function, batch size, dropout, and number of epochs. The values of the hyperparameters can be manually tuned or adjusted based on researchers' prior experience or intuition. This approach is a trial-and-error method which is inefficient. Efficient adjustment of hyperparameters can be achieved by using tuning model which can automatically modify the values of hyperparameters based on the prevailing performance of the DL models. Grid search, random search, and Bayesian optimization are the most prevalent tuning models for hyperparameter optimization for better performance.

Studies have developed sentiment analysis models and compared the performances of the models (Erik and Marie-Francine, 2008; Tan and Zhang, 2008; Prabowo and Thelwall, 2009; Anuj and Shubhamoy, 2012; Moraes et al., 2013; Neethu and Rajasree, 2013; Gautam and Yadav, 2014; Zhao et al., 2014; Medhat et al., 2014; Ortigosa et al., 2014; Selvi et al., 2015; Ghiassi et al., 2016; Jain and Dandannavar, 2016; Ahmad et al., 2017; Baid et al., 2017; Dhaoui et al., 2017; Mariel et al., 2018; Hasan et al., 2018; Kirilenko et al., 2018; Bajeh et al., 2018; Ali et al., 2019; Mohamed et al., 2019; Wu et al., 2019; Fitri et al., 2019; Krishna et al., 2019; Drus and Khalid, 2019; Garrick, 2020; Phan et al., 2020; García-Ordás et al., 2021; Liu et al., 2021; Abo et al., 2021; Daudert, 2021; AlGhamdi and Khatoon, 2021; Tran et al., 2021). However, the comparative studies do not report the empirical effect of hyperparameter optimization on the performance of the models. This study compares the performance of sentiment analysis models whose hyperparameters are optimized using tuning model against those that are not optimized. The random search tuning model is utilized to optimize the hyperparameters of the DL models considered in this study.

The remaining part of this paper is organized as follows. Section 2 presents a review of related works on sentiment analysis. The framework of this study is presented in section 3. Section 4 presents and discusses the results of the relative performances of the sentiment analysis models with and without hyperparameter tuning. The paper is concluded in section 5.

## 2. Related Works

Liu et al. (2021) proposed improving sentiment analysis accuracy with emoji embedding. Emojis, which contain hundreds of graphic symbols and are rapidly being used for expressing emotion in online chats, was incorporated as a new source of personalized expression. When constructing features for comparison with those directly derived from emoji label words, emojis were additionally translated into equivalent sentiment words. CEmo-LSTM is an improved emoji-embedding model based on long short-term memory. To investigate the sentiment evolution of online users during the COVID-19 pandemic, the CEmo-LSTM algorithm was deployed to a dataset collected from Weibo from the 1st of December 2019 to the 20th of March 2020. COVID-19 was found to have a significant

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023], pp. 139-159 https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

impact on individual moods and to create more passive emotions (e.g., horror and sadness). The accuracy of the experimental results is 95%. Because the hyper-parameter was not tuned, using a tuning model to tune the hyper-parameter could be considered an improvement in future study.

García-Ordás et al. (2021) presented a fully convolutional neural network to analyze sentiment in non-fixed length audios. A sentiment analysis approach was suggested that may handle audio of any length without being fixed a priori. Mel Frequency Cepstral and Mel Spectrogram As audio description approaches, coefficients are employed, and a fully convolutional neural network design is used as a classifier. Three datasets were used to test and train the model: EMODB, RAVDESS, and TESS. On the EMOBD, RAVDNESS, and TESS datasets, the model achieved an accuracy of 75.28 %, 92.71 %, and 99.03 %, respectively.

Abo et al. (2021) introduced the use of multi-criteria techniques for Arabic language sentiment classification for online product reviews in Saudi Arabia. Prominent DL and ML algorithms were empirically studied in this paper. The evaluation of the algorithms was based on recall, F-measure, precision, CPU time, classification error, area under the curve (AUC) and accuracy. SVM and the DL classifiers showed better performance, with precision of 85.30% and 83.87%, accuracy of 85.25% and 82.30%, f-measure of 86.81% and 83.87%, AUC of 0.93 and 0.90, and recall of 88.41% and 83.89% respectively. Other algorithms studied are K-nearest neighbors (K-NN), NB and Decision trees classifiers. The number of machine learning algorithms used in future work can be expanded, and combining the Arabic dialect dataset with the modern standard Arabic dataset will assist to generalize the results and improve this research. Daudert (2021) proposed a novel approach for fine-grained sentiment analysis using textual and relationship information for financial record. The proposed system enhances performance by up to 15% and 234% when compared to numerous baselines considered in the study. Investigating the proposed approach using more datasets was considered as a future work.

AlGhamdi and Khatoon (2021) compared the performance of three common ensemble approaches: Bagging, Adaboost, and Stacking for sentiment categorization prediction using five base learners: Decision tree, KNN, SVM, NB and Linear regression. Experiments were conducted on three different types of internet reviews: restaurant reviews, phone reviews, and movie reviews. As expected, the ensemble approaches outperformed the individual classifiers in general, with an average precision and recall of 0.83 and 0.82 respectively. Stacking (heterogeneous ensemble approach) was shown to be the best of the three ensemble methods, while bagging (homogeneous ensemble method) had the lowest performance. Tran et al. (2021) performed sentiment analysis on two separate movie review datasets using NB, Decision Tree, SVM, RNN, Voting and Blending. The experimental results of the study showed that notably the voting and RNN-based classification techniques had better performance.

Phan et al. (2020) proposed using the feature ensemble model to improve the capacity of sentiment analysis of tweets with fuzzy sentiment. The information linked to the word-type, semantic, lexicon, polarity sentiment and location of words is used to construct the twitter embedding. The ensemble model was created by combining data from five feature vectors: lexical, word-type, semantic sentiment polarity, and word placement in tweets containing fuzzy sentiment phrases. According to the findings of the study, the proposed method significantly improves the sentiment classification of retweets with fuzzy sentiment. Only tweets with a hazy sentiment were included in this study. Other elements such as slang and sarcasm could be considered in future work.

Garrick (2020) used machine learning to investigate the impact of gender and age on sentiment analysis using SVM, maximum entropy, long short-term memory and CNN. The dataset was generated by gathering book evaluations from Facebook users who were asked to fill out a questionnaire that would include questions about their reading tastes, as well as their different ages and gender. The generated data was separated into four groups on how users express their opinions. In all classifications, the over 50 age range beats all other age groups, according to the findings, with CNN and SVM classifiers having the maximum accuracy of 78%. Future work in delivering full examinations from multiple perspectives will benefit from the exploration of assessments in visual and audio form to identify emotions through voice and facial gestures.

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023], pp. 139-159 https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Mohamed et al. (2019) used DL models to analyse the sentiments of reviews of movie dataset and compared the performances of different deep learning networks. The study used Multilayer Perceptron (MLP) as a benchmark for other networks' results. IMDB dataset containing 50,000 movie review files with 50% positive ratings and 50% negative ratings was used for the analysis. CNN, LSTM, RNN, and a hybrid model incorporating LSTM and CNN were developed and tested. Word2Vec was used to pre-process the data before applying word embedding. According to the findings, the hybrid CNN LSTM model surpassed MLP and standalone CNN and LSTM networks with an accuracy of 89.2 %. CNN had 87.7% accurate, while MLP and LSTM had 86.74% and 86.64 % accurate respectively.

Iqbal et al. (2019) proposed a hybrid framework for sentiment analysis. The study proposed a new feature reduction strategy based on genetic algorithm (GA) that reduced the feature set size by up to 42% using this hybrid technique without impacting accuracy. When compared to more widely used latent semantic analysis (LSA) and principal component analysis (PCA) based feature reduction algorithms, the proposed hybrid framework performed better. It was observed that PCA had a 15.4 % increase in accuracy and LSA had a 40.2 % increase inaccuracy. To demonstrate the applicability of the GA-based designs, a new cross-disciplinary issue of geopolitics was offered as a study implementation for the sentiment analysis framework. The findings of the experiment showed that it was possible to accurately measure public attitudes and views on a variety of themes, including terrorism, global wars, and social issues. Extending this approach to cyber-intelligence could be a step forward in future work, allowing law enforcement authorities to generate suggestions based on user feedback.

Wu et al. (2019) conducted sentiment analysis of Chinese microblogs using various sentiment dictionaries and a set of semantic rules. An algorithm was proposed for determining Chinese microblog sentiment from difficult phrases to clauses, clauses to words, and finally words to emoji. A new word feeling dictionary for Chinese microblogs was created. The coverage of sentiment words is increased by using multiple sentiment dictionaries. Simultaneously, semantic rule sets between Chinese microblog texts were investigated. Sentence analysis rules and sentence pattern analysis rules were incorporated into the analysis to enhance the accuracy. The results of the experiments showed that this strategy significantly improves sentiment analysis of Chinese microblogs.

Ali et al. (2019) proposed a multinomial Naive Bayes (MNB) classification model for sentiment analysis of movie reviews. Each review includes a text note as well as a numerical score between 0 and 100 (inclusive). The accuracy of the experimental results is 90%. In the future studies, training the algorithm with more datasets from other sources could be regarded as an upgrade. Fitri et al. (2019) studied NB, random forest and decision tree algorithms for analyzing sentiment of Twitter using the issue of Anti-LGBT campaign in Indonesia as a case study. The accuracy of the experimental result is 86.43%. The result of the analysis showed that Twitter users in Indonesia post more neutral remarks.

Krishna et al. (2019) employed SVM, KNN, RF and NB for the sentiment analysis of restaurant reviews. The results of the analysis showed that the SVM classifier had the highest accuracy of 94.56 %. The study recommended that additional machine learning algorithms could be investigated in future studies. Ali et al. (2019) investigated various machine learning techniques in order to determine their importance and to pique interest in this field of research. At the end of the fall 2018 semester, Google Survey Forms was used to collect student feedback from both public and private universities in Karachi. The dataset contains useful information about teaching and learning quality. Statistical Analysis/Methods: SVM, RF, MNB, MLP and Stochastic Gradient Decent Classifiers were modeled in the study. The MNB and MLP have the maximum accuracy.

Drus and Khalid (2019) proposed sentiment analysis in social media using a hybrid approach of Lexicon based method (SentiWordnet and TF-IDF) and machine learning method (Naïve Bayes and SVM). The following trusted and credible databases were used to conduct a systematic review of papers published between 2014 and 2019: ACM, Emerald Insight, IEEE Xplore, Science Direct, and Scopus. 24 out of 77 papers were picked from the review process after an initial and in-depth screening. The results showed that lexicon-based technique were been initially used for sentiment analaysis. Machine learning-based methods are better suited to larger data because they require more

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

time and data to train. Combining lexical and machine learning methods to increase the quality and accuracy of the outcome is recommended. Secondly, Twitter was seen to be the most prevalent social media site from which sentiment analysis data is extracted. That is, majority of the reviewed papers used Twitter as their source of social media data. This is due to Twitter's wide availability, accessibility, and diversity of content. Every day, millions of tweets are sent out on practically any subject. This suggests that social media is rapidly evolving into a valuable source of knowledge. Other social media platforms, such as blogs, WordPress, YouTube, and others, receive less attention. The content of each social media platform may differ, so it's worth looking into alternative options that could lead to new insights and discoveries. Finally, we saw how sentiment analysis may be used in social media. Sentiment analysis can be used to improve business quality and strategy, predict election results, monitor disease outbreaks, raise awareness about the need of data security, improve perceptions of a particular sport, and improve catastrophe location and reaction, among other things. This demonstrates how sentiment analysis aids decision-making by helping to understand people's perceptions. Further study is needed to develop a universal sentiment analysis model that can be used to a range of data formats, as well as to look into other potential social networking sites for gathering user feedback and expanding the scope of sentiment analysis applications.

Mariel et al. (2018) compared deep learning neural network algorithm with Naive Bayes (NB) and SVM on Indonesian text. Different combinations of text preprocessing and feature extraction approaches were used for the modeling, taking into account the content of sentiment in the text. The results of the deep learning neural network model were compared to that of NB and SVM. The Indonesian text data contain both balanced and unbalanced sentiment composition. As expected, the deep learning neural network model outperformed the NB and SVM models in terms of F1 score which is the harmonic mean of precision and recall. Tuning the hyperparameter of deep learning with a tuning model could be regarded an improvement in future research.

Kirilenko et al. (2018) studied automated sentiment analysis in tourism; the study assesses the applicability of various types of automated classifiers for tourism, hospitality, and marketing studies. The study compared automated analysis with that of human raters. While commonly used performance indices suggest that machine learning algorithms perform similarly to human raters on easier-to-classify data sets, other performance metrics, such as Cohen's kappa, show that machine learning outcomes are still inferior to manual processing. Automated analysis performs worse than human raters on more difficult and noisy data sets. Bajeh et al. (2018) conducted a comparative analysis of the performance of NB, K-NN and SVM for sentiment analysis. Two large textual datasets describing the opinions of customers on airline services and the 2015 Superbowl show were collected from Tweeter for the comparative analysis. The Scikit-Learn machine learning library of Python programming language was used for the implementation of the models. Their performance were measured in terms of precision, recall and accuracy. The results of the study showed that K-NN has the best performance in terms of accuracy with measures of74.5% and 60.3% in the airline and Superbowl datasets respectively. NB has the best performance in terms of precision and recall with a measure of 92% and 100% respectively in the airline dataset, and 82% and 96% respectively in the Superbowl dataset.

Hasan et al. (2018) proposed using machine learning-based analysis on sentiment for twitter accounts. In this work, supervised machine-learning algorithms such as NB and SVM were used to compare sentiment analysis techniques in the analysis of political opinions. In the first step of sentiment analysis, tweets from multiple sentiment analyzers were reviewed to see how accurate they were. The major purpose of evaluating the accuracies of several sentiment analyzers like Text Blob, SentiWordNet, and WSD was to see how accurate they were. There were 100,000 tweets aggregated from public accounts. Only 6250 tweets are left after pre-processing. Among the three sentiment analyses evaluated in this study, TextBlob had the highest rate of positive sentiment tweets, with 3380 tweets and a percentage of 54.08%. SentiWordNet assigned a negative sentiment rate of 48.86 % to 3054. To validate the sentiments (positive, neutral, negative) of tweets received from several sentiment analyzers, NB was used on a 1690 training dataset of 845 favorable and 845 ‘ negative assessed by each sentiment analyzer. The test set, on the other hand, comprised 400 tweets. W-WSD had the highest accuracy rate for predicting election opinions when tested with the NB, with 316 correct examples and a 75% accuracy rate. TextBlob was found to be the most accurate, with 304 correct instances and a 76% accuracy. In comparison to W-WSD and TextBlob, SentiWordNet's accuracy was

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023], pp. 139-159 https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

quite low, yielding only 219 accurate occurrences with a 54.75% accuracy. When comparing W-WSD with TextBlob, the results show that TextBlob has the highest accuracy, and Textblob and W-WSD are far superior to the SentiWordNet technique for understanding election feelings and providing more accurate forecasts than the SentiWordNet approach. To take the study further, future research could look for trends of political parties based on Twitter ratings.

Ahmad et al. (2017) reviewed several studies that used ML algorithms for sentiment analysis using variety of datasets including movie and product evaluations. It was observed that SVM did better in terms of efficiency and accuracy. The average and relative accuracy of all machine learning-based sentiment classification approaches evaluated is summarized Table 1.

Table 1: Tools, Features and their accuracy on different datasets (Ahmad et al., 2017)

| SR# | Author Name | Publication Year | Tool Name | Features | Accuracy | | |
|---|---|---|---|---|---|---|---|
| | | | | | Movie | Product | Average |
| 1 | K. Nigam, J. LaÆerty [14] | 1999 | Maximum Ent | - Feature based Model<br>- Uses, bigrams and phrases | - | - | 72.60% |
| 2 | L. Breiman [17] | 2001 | Random Forest | - Ensemble learning method<br>- Uses Classification trees and tree predictors<br>- comprised of tree structured classifiers | - | - | 88.39% |
| 3 | Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S. [1] | 2012 | SASA | - Based on sentistrength ver. 2.0<br>- Used by Amazon AMT<br>- Uses IBM's InfoSphere for speed and performance. | - | - | 64.90% |
| 4 | P.K Singh and M.Shahid Hussain [20] | 2014 | MLP | - Robust & non-linear neural network model<br>- Used as universal function approximator<br>- it have one hidden layer and multiple non-linear units | 81.05% | 79.27% | - |
| 5 | P.K Singh and M.Shahid Hussain [20] | 2014 | Naïve Bayes | - Supervised classifier for binary classification<br>- Based on Baye's Theorem<br>- Useful for large datasets | 75.50% | 62.50% | - |
| 6 | C. Cortes and V. Vapnik [24] | 1995 | SVM | - based on supervised learning model<br>- can handle linear separation on high dimensional non-linear input data using an appropriate kernel<br>- multiple derivatives and extensions are available | 81.15% | 79.40% | - |

Baid et al. (2017) studied the use of K-NN, NB and RF machine learning techniques for sentiment analysis of movie reviews. The dataset used was from the Internet Movie Database, 2000 IMDb-archived user-created movie reviews consisting of 1000 positive and 1000 negative reviews. The data were transformed to arff format, and then pre-processed using the WEKA tool. NB had the best results with an accuracy of 81.45%. RF had an accuracy of 78.65%, and K-NN had an accuracy of 55.30%.

Dhaoui et al. (2017) proposed social media sentiment analysis using a hybrid approach of lexicon and machine learning. A sample of 850 consumer comments from 83 Facebook pages were used to evaluate and compare lexicon-based and machine learning methodologies to sentiment analysis, as well as their combination, using the LIWC2015 vocabulary and the RTextTools deep learning package. When applying machine learning and lexicon-based approaches to sentiment analysis, the results show that F scores for positive and statistically significant classification (F 0.77 and 0.78, respectively) and negative sentiment classification (F 0.45 and 0.47, respectively) are reasonably

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023], pp. 139-159 https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

close. The findings also show that both techniques are more accurate at classifying positive valence than negative valence. When applied to social media interactions, these findings show that existing sentiment analysis algorithms are good at predicting positive valence but not so good at predicting negative valence. More research is needed to increase the performance on classifying negative sentiment. The combined technique should be used for other forms of CGCs on social networks, such as tweets. Furthermore, while the results of combining the two algorithms are promising for positive sentiment analysis, additional study is needed to improve negative sentiment analysis accuracy.

Ghiassi et al. (2016) proposed using supervised feature engineering and the dynamic architecture for artificial neural networks for targeted twitter sentiment analysis (TSA) for brands. In this study, supervised feature engineering and DAN2, a machine-learning model with the sensitivity to distinguish mild sentiment expressions, were used to propose a tailored approach to TSA for brands. The solution addresses the challenges that come with Twitter's unique vocabulary, as well as the memory of mild sentiment expressions, all of which are important to businesses and brand managers. The utility of the proposed technique was demonstrated using two Twitter data sets relating to the Starbucks and Governor Chris Christie brands. These feature representations were merged with DAN2 to categorize tweet sentiment. A comparison of the proposed method's performance in three-class and five-class tweet sentiment classification with the results of several competitors, including two cutting-edge TSA systems from the academic (Sentiment140) and commercial (Repustate) worlds. SVM was used to isolate and analyze DAN2's performance using the same Twitter feature representations as another popular machine-learning model. The results demonstrated that the suggested supervised feature engineering technique for brands' tweet feature representations outperformed the state-of-the-art Sentiment140 and Repustate TSA systems by a significant margin, with classification precisions and recalls frequently surpassing 80%. The research has various limitations and future directions. Even though the Twitter datasets are large enough to train a machine-learning classifier, the brand-related Twitter datasets are small because manually annotating tweets to create gold standard sentiment class labels for classifier training and evaluation is time-consuming and laborious. A further study of the additional brand-related cases and Twitter datasets could be done. Also, to improve tweet feature representations and performance in TSA, a transition to supervised feature engineering may be contemplated in the future.

Jain and Dandannavar (2016) proposed sentiment analysis using machine learning algorithms on twitter data. The proposed system employs machine learning methods such as Decision Trees Multinomial and Naive Bayes for sentiment analysis. This study examined datasets of various sizes and domains to show that the suggested framework works with data of all sizes and domains. For each domain (IT Industry (Apple), Bank (ICICI), Telecom (BSNL), 200 to 4000 tweets were collected, and the relevant findings were examined. 70% of each dataset is used to train the classifier, and the remaining 30% used for testing. The results of the analysis showed that the Decision Tree performs exceptionally well, with 100% accuracy, precision, recall, and F1-Score.

Selvi et al. (2015) conducted a comparative study of feature selection and machine learning algorithms for sentiment classification of movie dataset. This study compares the accuracy of six feature selection mechanisms (CHI, OneR, Relief-F, SAE, IG and GR) with five different machine learning classifiers: DT, K-NN, ME, SVM and NB on the movie review dataset. With an accuracy of 92.95%, 95.3%, 92.95%, and 95.6%, NB outperformed other classifiers for the feature selection methods GR, CHI, SAE and IG respectively. ME outperforms all other classifier algorithms for the feature selection methods RF and OneR, with accuracy of 70.65% and 88.6% respectively. NB performed better since it is based on the likelihood of phrases occurring in the text. NB, ME, and SVM fared better in the range of 2,500–3,500 features, and K-NN and DT performed better for features exceeding 2,000. The RF feature selection method was found to be inaccurate for the movie review dataset. When using the GR, IG, CHI, OneR, and SAE feature selection methods, the NB classifier is more accurate. It was revealed that NB with SAE and GR feature selection method works well in movie reviews. Gautam and Yadav (2014) did sentiment analysis using ME, NB and SVM, and WorldNet which classify the sentence and product reviews sentiment analysis of twitter data. A data set of 19340 instances of which 18340 instances were utilized for training and 1000 instances were used for testing The NB technique outperformed ME and SVM. When WorldNet is combined with the above approach, the accuracy increases to 89.9% from 88.2%. To enhance the features vector-related sentence recognition process, it is

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

recommended that the training data set be enlarged in the future, and that WorldNet be utilized to summarize the reviews. It may provide a better visual representation of the content, which will be beneficial to the users.

Zhao et al. (2014) proposed sentiment analysis on news comments based on supervised learning method. Four feature selection approaches (DF, MI, IG, CHI), three feature representations (TF, Presence, TF-IDF), and five learning methods (ME, Winnow, NB, C4.5, SVM) were utilized to analyze the sentiment of Chinese news comments. The goal was to improve how news comments were classified in terms of mood. Document frequency (DF), IG, and CHI were found to be effective feature selection strategies for all classifiers. Different accuracies are achieved by classifiers with DF, IG, and CHI; nonetheless, the difference is minor. The superior feature selection approach is CHI, which is a comprehensive examination. MI has the worst effect when compared to DF, IG, and CHI. The results showed that MI is not suited for assessing the sentiment of news comments. For all classifiers with DF, IG, and CHI, TF is somewhat better than the other two feature representation methods, and the difference between the three feature weighting methods is minor. As a result, for three feature representation approaches, the textual vectors of the same text is comparable. The results of the experiment reveal that machine learning is quite good at classifying the sentiment of news comments. When compared to the other five classifiers, NB, SVM, and ME are the top classifiers for sentiment classification of news comments. When integrating feature selection methods, feature representation methods, and different classifiers to investigate the impact of sentiment classification on news comments, it was observed that feature selection methods and feature representation methods have just a maginal impact. Increasing the accuracy of sentiment classification by a substantial margin necessitates strengthening the classifiers. Analysis of binary sentiment classification, multi-level sentiment classification of other short articles, and studying ways to improve classifiers to increase classification accuracy and balance could be further studied.

Medhat et al. (2014) surveyed studies on sentiment analysis algorithm and their applications. A total of 54 were reviewed. WordNet, which is available in languages other than English, is the most used lexicon source. Many natural languages still require the creation of resources, which are employed in sentiment analysis. In recent years, information from microblogs, blogs, and forums, as well as news sources have become increasingly popular in South Africa. This type of media information is extremely useful in expressing people's feelings or opinions about products or some topic of interest. The study showed that IMDB dataset is prominent in sentiment analysis for measuring the performances of algorithms. In many applications, it's critical to consider the text's context as well as the user's preferences. As a result, more research on context-based sentiment analysis is required.

Ortigosa et al. (2014) proposed novel approach for Facebook sentiment analysis and investigated its applicability to e-learning. The approach gathers information about users' sentiment polarity (positive, neutral, or negative) as indicated in their messages, models people' regular sentiment polarity, and detects large emotional shifts. SentBuk, a Facebook application also introduced in this paper, was used to achieve these features. SentBuk collects messages sent by Facebook users and categorizes them based on their polarity, presenting the results to the users in an interactive interface. It also has features such as emotional change detection, detecting friends' emotions, user classification based on their communications, and statistics, among others. SentBuk's categorization algorithm takes a hybrid approach, combining lexical-based and machine-learning techniques. The findings obtained using this method demonstrated that a high accuracy of 83.27% in sentiment analysis on Facebook is possible.

Neethu and Rajasree (2013) conducted Twitter sentiment analysis using machine learning techniques. After data preprocessing, an efficient feature vector was constructed by doing feature extraction in two steps. The initial step extracted and added Twitter-specific features to the feature vector. Following that, the features were removed from tweets, and feature extraction was performed as if it were normal text. These qualities are reflected in the feature vector as well. To test the feature vector's classification accuracy, various classifiers such as SVM, NB, Ensembles, and Maximum Entropy are used. For the new feature vector, all these classifiers have nearly identical accuracy. This feature vector worked very well on electronic products sentiment analysis.

Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023], pp. 139-159 https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Moraes et al. (2013) conducted an empirical comparison of ANN and SVM for document-level sentiment classification. Experiments show that SVM is less influenced by noisy phrases than ANN as the data asymmetry increases. As expected, a neural network's training period is typically substantially longer than that of an SVM. If the issue is running time, ANN may be preferable since SVM often results in a large number of support vectors and as a result, a running time that develops much faster than that of ANN. IG, a computationally inexpensive feature selection strategy, was employed to lower the computing cost of both ANN and SVM without impacting the classification accuracy. With a growing number of input terms, the results showed that there are some thresholds above which there will be no gain in accuracy. According to the findings, IG makes ANN training practicable in a bag-of-words approach and helps to cut SVM running time.

Anuj and Shubhamoy (2012) suggested feature extraction and machine learning techniques for sentiment analysis on a dataset of online movie reviews. The study combined five most commonly used feature selection technique in data mining research: DF, information gain (IG), gain ratio (GR), CHI, and Relief-F, with seven ML classification methods: Maximum Entropy (ME), Decision Tree, NB, SVM, K-NN, Winnow, and Adaboost. The findings showed that feature selection techniques improved the performance of sentiment classification results. GR surpasses other strategies for sentimental feature selection, and SVM surpassed other ML approaches.

Prabowo and Thelwall (2009) proposed a combined approach for sentiment analysis. In this paper, supervised machine learning and rule-based classification are combined to create a novel method. Film reviews, product reviews, and Myspace comments were all used to test this strategy. In terms of macro- and micro-averaged F1 score which is the harmonic mean of precision and recall, the findings showed that utilizing a hybrid classification enhances classification efficacy. A semi-automated, complimentary strategy was also presented, in which each classifier contributes to the effectiveness of other classifiers. The second rule set was created by substituting each proper word found within each phrase to construct a set of antecedents, and then assigning sentiment to each antecedent (the formation of a set of rules). The underlying assumption was that each antecedent's attitude was equivalent to the feeling ascribed to the pre-classified document in which the antecedent was discovered. Then a rule-based classifier (RBC) was devised, which used this second rule set to classify a collection of document. It is possible that the antecedent expresses a feeling that differs from the sentiment expressed in the related document. As a result, a sentiment analysis tool (SAT) was used.

Erik and Marie-Francine (2008) studied the use of the ML techniques: MNB, SVM, and ME to analyzing sentiment in multilingual web texts in blog, review, and forum writings written in English, Dutch, and French. The training was based on a set of example sentences or utterances that were manually labelled as positive, negative, or neutral. The result of the study showed that in classifying English, Dutch, and French Web data according to sentiment, unigram features augmented with a small number of language-specific features produce accurate results of around 83%, 70%, and 68% respectively, and slightly improve a baseline classification that only uses unigrams. The best results were observed with a MNB classifier for English, an SVM for Dutch, and a ME classifier for French. It was claimed that because the language in the corpora used for the analysis is so diverse, many of the errors can be traced back to a lack of training instances. The use of active learning techniques to improve annotation was investigated. The issue can be solved at the preprocessing level by fine adjustment, which is required such that the mapping's target words are less specific but not too broad. Improvements or enhancements to the machine learning framework were also made, as well as a customized example of a cascaded architecture in the trials. Many additional algorithm configurations, feature extractions, and training set compositions can be used in the quest for the best classification models and consequently the best outcomes.

Tan and Zhang (2008) conducted empirical research on sentiment analysis on Chinese document. The study used a Chinese emotion corpus of 1021 documents, four feature selection approaches: IG, MI, DF, and five machine learning algorithms: K-NN, centroid classifier, NB, SVM and winnow classifier. The experimental results show that IG is the most effective at selecting sentimental phrases, whereas SVM is the most effective for sentiment classification. Furthermore, sentiment classifiers are found to be highly dependent on domains or subjects. As there was a lack of publicly available Chinese sentiment corpus for evaluating sentiment analysis tools, a Chinese

Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023],
pp. 139-159 https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

sentiment corpus was created. It was referred to as "Chn-SentiCorp" for the sake of convenience. There are 1021 documents in all, divided into three categories: education, film, and home. There were 507 items linked to education, 266 materials related to movies, and 248 documents related to houses. In each domain category, there are documents that are both positive and bad. The results of the experiments showed that IG outperforms all other learning approaches. It has an average MicroF1 of 0.8883, which is 1% higher than CHI (0.8737), 2% higher than DF (0.8644), and 8% higher than MI (0.8086).

Figure 1 presents a taxonomy of the reviewed studies.

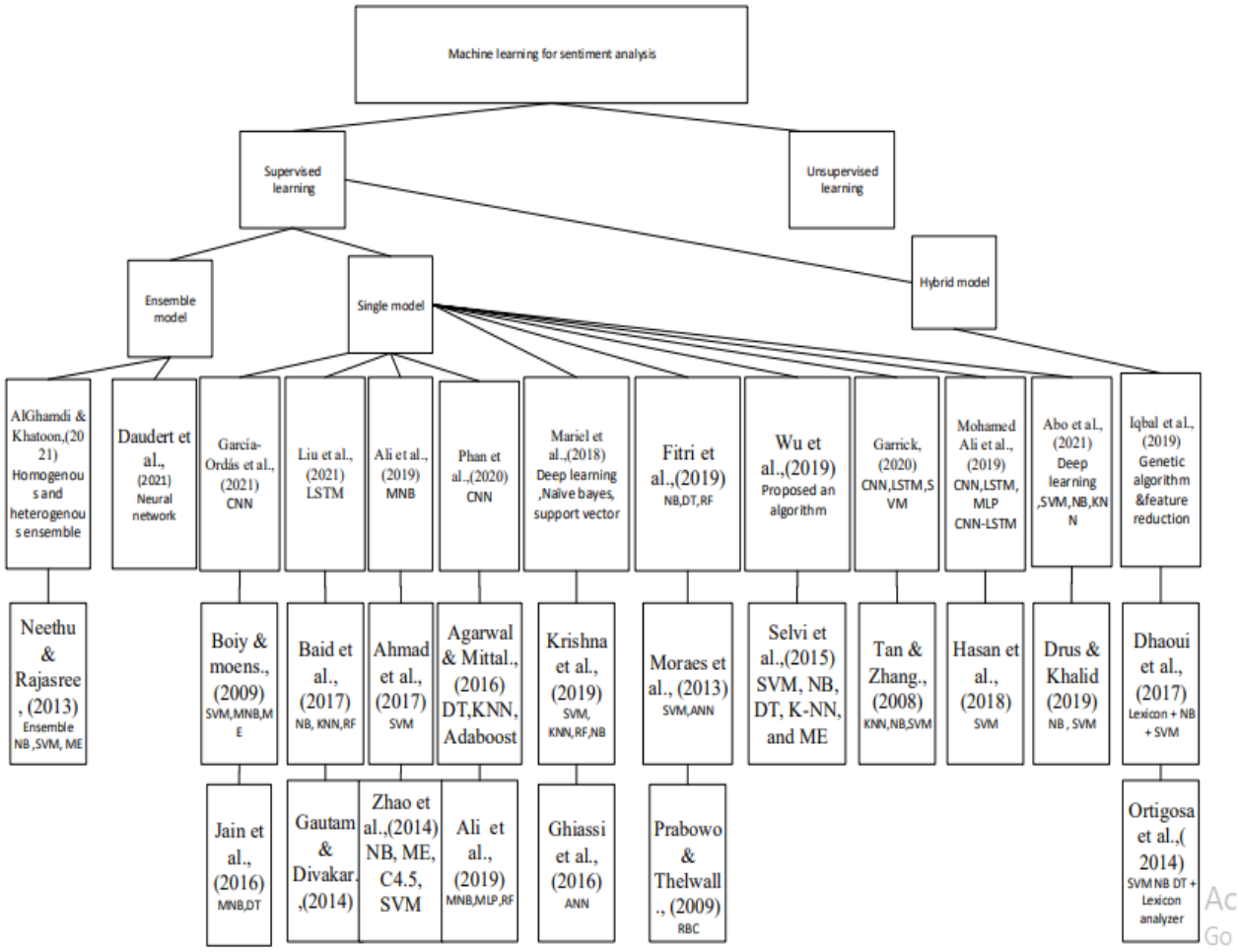


Figure 1: A Taxonomy of the literature on Sentiment Analysis

The reviewed studies showed that although several studies have been conducted to propose different sentiment analysis models with the view of improving performance, none reported or thoroughly investigated the impact of hyperparameter tuning on the performance of sentiment analysis models. Thus, there is a need to empirically investigate the effect of optimizing hyperparameters on performance. This study investigates the effect of hyperparameter optimization on the performance of DL approaches to sentiment analysis. MLP, LSTM, CNN, and CNN/LSTM hybrid are the deep learning algorithms investigated in this study.

## 3. Methodology

This section presents the research framework and methods used to investigate the impact of hyperparameter tuning on the performance of DL models. Figure 2 depicts the research framework used in this study.

Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023],
pp. 139-159 https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

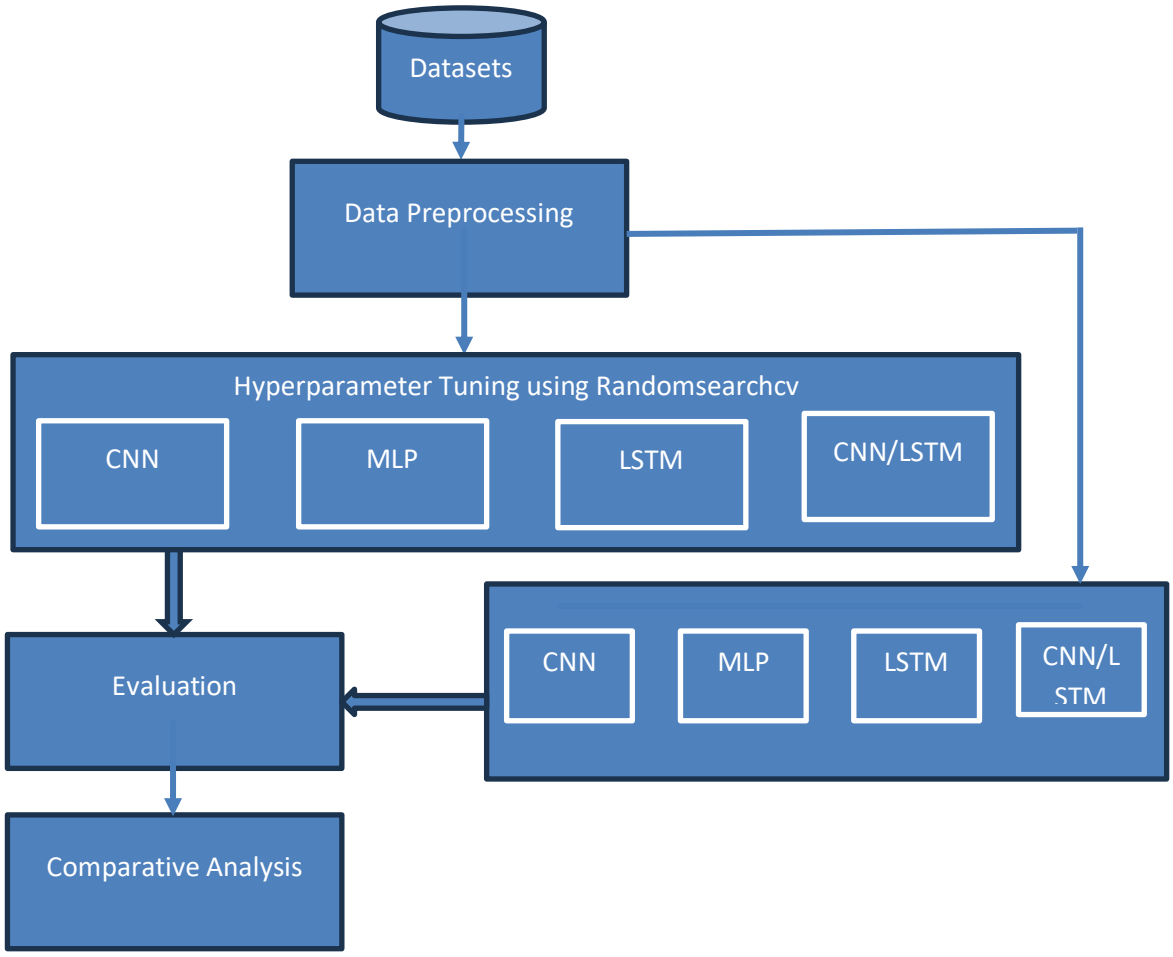


**Figure 2 Research framework**

The study framework involves the preprocessing of the datasets to putting them in a form suitable for training the DL models. This include removing duplicates and unwanted character and or words, tokenization, normalization, word embedding and sequencing. Thereafter, the preprocessed data is used to train two sets of the DL models (CNN, MLP, LSTM and CNN/LSTM). One set involves hyperparameter tunning using random search technique (randomsearchcv in Python), while the other set of DL models are not optimized but directly trained using the preprocessed data. The preprocessed data is split into training set and testing set using the 70/30 split respectively. The trained models are evaluated using the classification performance metrics, precision, recall and accuracy. The results of the evaluation are compared with each other and also with a related study reported in the literature.

### 3.1 Datasets

Three sentiment analysis datasets from Kaggle will be used for testing and training the selected deep learning algorithms in this study: IMDB movie reviews, stock market sentiment dataset, and Twitter US airline sentiment dataset.

The IMDB movie review comprises two columns described in Table 2. The dataset contains 50,000 instances of viewers' opinions about a movie.

Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023], pp. 139-159 https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

**Table 2: IMDB Movie review dataset**

| Attributes | Description |
|---|---|
| Review | People's opinion about movies |
| Sentiment | Polarity of the opinion or review (Label) |

The stock market sentiment dataset is a collection of people's opinions about the stock market. It has 5792 instances and 2 columns described in Table 3.

**Table 3: Stock Trading Dataset**

| Attributes | Description |
|---|---|
| Text | People's opinion about stocks |
| Sentiment | Polarity of the opinion (Label) |

Twitter US airline sentiment dataset contains the expressions of US airline travelers' opinions on Twitter about the services render by US airline. The dataset has 15 columns as described in Table 4, and 14,640 instances of travelers' review comments.

**Table 4: Twitter US Airline Sentiment dataset**

| Attributes | Description |
|---|---|
| tweet_id | Passenger Twitter id |
| airline_sentiment | Polarity of sentiment (Label) |
| airline_sentiment_confidence | Degree of confidence in sentiment ranging from 0 to 1 |
| negativereason | Negative reasons for the opinion |
| negativereason_confidence | Degree of confidence in the negative reason ranging from 0-1. |
| Airline | Name of airline |
| Name | Name of passenger |
| retweet_count | 0 when a passenger did not retweet. |
| Text | Tweet |
| tweet_coord | Tweet coordinate |
| tweet_created | Tweet |

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

| | |
|---|---|
| tweet_location | Location where passengers make tweets |
| user_timezone | User time zone |
| Zone | Time zine |

### 3.2 Data Preprocessing

The performance of machine learning models is partly a function of the goodness of the dataset used to train the models. That is, similar to the computer system, it is garbage in garbage out. Thus, data processing is required to put it in a suitable form for modeling. This preprocessing is because data is often noisy and contains missing or erroneous information. For instance, in sentiment datasets, individuals express their opinions about products and/or services in a variety of ways. This can put the data in an inconsistent form not suitable for modeling. Inconsistency and duplication are common problems with polarized raw data. Preprocessing is also concerned with removing redundant tokens and character that do not contribute meaningfully to the data. For instance, text such as "That cloth is Beauuuutifull #," can becomes "cloth Beautiful", and "Juliet is Noww Hardworkingg" can becomes "Juliet now hardworking." after preprocessing. Some datasets must be transformed to a machine learning algorithm-friendly format.

Tokenization, normalization, word embedding, and sequencing are the preprocessing tasks that are performed on the dataset. Tokenization is the partitioning of data into a large vector of words. This entails breaking the document into sections, adding new lines, and tabs, and then saving it again. The split () method on the loaded string can be used to accomplish this in Python. Normalization of the text is done by converting all text to lower case, punctuation marks are removed, and numbers are replaced with their equivalent words. During word embedding, each word is represented as a set of real vectors in a predetermined vector space. After that, each word is assigned to a single vector. In this study, the Keras library's word2vec embedding is used with a vector size of 32. Sequencing converts the numeric array from the word2Vec embedding into a digital vector, which is then passed to the proposed deep learning model.

### 3.3 Hyperparameter Optimization

With the aim of get the best results from deep learning models, the hyperparameters of the models are adjusted with a tuning model. The purpose of hyperparameter tuning is to discover a set of hyperparameter settings that gives a deep learning model the optimum performance. There is no fixed rule for establishing these hyperparameters because the settings for different tasks vary, however there is a set of values at which the model performs at optimum; this set of values is what is determined by the tuning model. The hyperparameters tuned in this study are the **number of epochs** which was varied between 40, 60, and 100, and **batch size** which was varied between 32, 64, and 128. Randomsearchcv was used to tune the system. A grid of hyperparameter values was entered to implement Randomsearchcv. The model was then trained using a random mix of hyperparameter values. This determines how many parameter combinations are attempted.

### 3.4 Model Implementation

The DL models considered are implemented using the Python programming environment. Python is known for its readability and simplicity. It provides sophisticated DL frameworks and libraries such as TensorFlow and Keras. In this study, anaconda IDE, Keras framework, pandas, and word2vec are used for the implementation of the DL models and the analysis of the results.

### 3.5 Performance Evaluation

Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023],
pp. 139-159 https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

The performance of the DL models are measured in terms of precision, recall and accuracy. Equation 1, 2 and 3 presents the mathematical expression for determining these performance measures.

$$precision = \frac{TP}{TP+FP} \quad (1)$$

$$recall = \frac{TP}{TP+FN} \quad (2)$$

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Where *TP*, *TN*, *FP* and *FN* are defined as follows with respect to this study:

*True Positive (TP):* the number of positive sentiments correctly predicted as positive.
*True Negative (TN)*: the number of negative sentiments correctly predicted as negative.
*False Positive (FP)*: the number of negative sentiments wrongly predicted as positive.
*False Negative (FN):* the number of positive sentiments wrongly predicted as negative.
Table 5 further presents these metrics in the form of confusion matrix.

Table 5: Confusion Matrix

| | | Predicted sentiment | |
|---|---|---|---|
| | | +ve sentiment | -ve sentiment |
| Actual sentiment | +ve sentiment | ***TP*** | ***FN*** |
| | -ve sentiment | ***FP*** | ***TN*** |

## 4. Result and Discussion

This section presents and discusses the result of the data preprocessing and analysis carried out in this study as described in the previous section.

### 4.1 Data Preprocessing and Hyperparameter optimization

Figure 3 depict a snapshot showing the result of the data preprocessing. Each of the datasets was tokenized, normalized, word embedded and sequenced. The result of the preprocessing of each dataset is a digital vector which serves as input to the training of the DL models. The hyperparameter tuning is carried out using the random search technique. Figure 4 depicts the implementation of the hyperparameter tuning process. The number of epochs and batch size were the hyperparameters that were optimized. As shown in the figure, the number of epochs was calculated using 40, 60, and 100, while the batch size was calculated using 32, 64, and 128. One parameter at a time, the modeling chooses a value and observe the performance of the model as it is being trained. The final values of the two parameters are the values that yielded the best training result. For the hyperparameters, epoch = [40, 60,100] and batch size = [32, 64,128], a distribution was first created and then Randomsearchcv iteratively make selections from the distribution. Randomsearchcv does not consider the entire distribution but rather takes a random sample of it. In this study, the number of iterations was set to 5.

Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023],
pp. 139-159 https://doi.org/10.56471/

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

```
tokenizer = Tokenizer(num_words=max_words)

tokenizer.fit_on_texts(X_train)

X_train_seq = tokenizer.texts_to_sequences(X_train)
X_train_seq = sequence.pad_sequences(X_train_seq, maxlen=max_len)
X_train_seq
```

```
[24]: array([[   0,    0,    0, ...,  147, 3977,  335],
       [   0,    0,    0, ...,  809,  147,  888],
       [   0,    0,    0, ...,   21,  205, 1706],
       ...,
       [   0,    0,    0, ...,  178,  540, 2381],
       [   0,    0,    0, ...,  715,   20,  976],
       [   0,    0,    0, ..., 3778,    5,    1]], dtype=int32)
```

Figure 3: Result of Data preprocessing

```
batch_size = [32,64,128]
epochs = [40,60,100]


param_grid = dict(batch_size=batch_size, epochs=epochs)

model2 = KerasClassifier(build_fn=get_cnn_model, verbose=1)

#grid = RandomizedSearchCV(estimator=model2,param_grid=param_grid, n_iter=5
Random=RandomizedSearchCV(estimator=model2, cv=5, param_distributions=param
Random_result = Random.fit(X_train_seq,Y_train)
```

Figure 4: Implementation of hyperparameter tuning for optimal performance.

### 4.2 Results of unoptimized DL models

The results of deep learning models without hyperparameter adjustment are presented in this section. Table 6 shows the results for the Twitter US airline sentiment dataset. CNN-LSTM has the highest accuracy of 89.85 %, followed by CNN with an accuracy of 87.98%.

**Table 6: The performances of the DL models on Twitter US Airline Dataset**

| S/N | Metrics | CNN-LSTM | CNN | LSTM | MLP |
|---|---|---|---|---|---|
| 1 | Accuracy | 89.85% | 87.98% | 86.99% | 87.01 |
| 2 | Precision | 0.8 | 0.85 | 0.75 | 0.79 |
| 3 | Recall | 0.90 | 0.82 | 0.60 | 0.70 |

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023],
pp. 139-159 https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Table 7 presents the results of the models on the stock market dataset. Also, the result shows that CNN-LSTM has the highest accuracy when compared to other deep learning models.

**Table 7: The performances of the DL models on Stock Market Dataset**

| S/N | Metrics | CNN-LSTM | CNN | LSTM | MLP |
|---|---|---|---|---|---|
| 1 | Accuracy | 89.10 | 87.60% | 86.54% | 86.35% |
| 2 | Precision | 0.79 | 0.78 | 0.73 | 0.79 |
| 3 | Recall | 0.77 | 0.78 | 0.68 | 0.77 |

Table 8 shows the result of DL models on the IMDB movie review dataset. Again, the CNN-LSTM has the highest accuracy which is followed by CNN.

**Table 8: The performances of the DL models on IMDB Movie Review Dataset**

| S/N | Metrics | CNN-LSTM | CNN | LSTM | MLP |
|---|---|---|---|---|---|
| 1 | Accuracy | 90.19 | 91.01% | 86.99% | 85.98% |
| 2 | Precision | 0.99 | 0.84 | 0.68 | 0.72 |
| 3 | Recall | 0.81 | 0.81 | 0.71 | 0.69 |

### 4.3 Results of optimized DL models

The results of deep learning models with hyperparameter adjustment are presented in this section. Table 9 shows the result of the optimized models on the Twitter US airline dataset. Once again, with an accuracy of 90.92%, precision of 0.89, and recall of 0.84, the CNN-LSTM model outperformed other models.

**Table 9: The performances of the optimized DL models on Twitter US Airline Dataset**

| S/N | Metrics | CNN-LSTM | CNN | LSTM | MLP |
|---|---|---|---|---|---|
| 1 | Accuracy | 90.92% | 89.99% | 87.95% | 88.05% |
| 2 | Precision | 0.899 | 0.845 | 0.740 | 0.811 |
| 3 | Recall | 0.840 | 0.821 | 0.619 | 0.802 |

Table 10 is the performance results of the optimized models on the stock market dataset. Like the results in Table 9, CNN-LSTM also outperformed other models except in recall where CNN outperformed the CNN-LSTM and other models. Similar to the results in Table 10, CNN-LSTM outperformed other models in terms of accuracy and precision, but CNN outperformed it and other models in terms of recall.

Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023], pp. 139-159 https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

**Table 10: The performances of the optimized DL models on Stock Market Dataset**

| S/N | Metrics | CNN-LSTM | CNN | LSTM | MLP |
|---|---|---|---|---|---|
| 1 | Accuracy | 91.02% | 88.97% | 87.85% | 88.01% |
| 2 | Precision | 0.806 | 0.801 | 0.740 | 0.802 |
| 3 | Recall | 0.78 | 0.805 | 0.689 | 0.784 |

Lastly, Table 11 presents the result of the optimized models on the IMDB movie review dataset. Also, like the results in Tables 9 and 10, CNN-LSTM outperformed other models in terms of accuracy and precision. It was marginally outperformed by CNN in terms of recall.

**Table 11: The performances of the optimized DL models on IMDB Movie Review Dataset**

| S/N | Metrics | CNN-LSTM | CNN | LSTM | MLP |
|---|---|---|---|---|---|
| 1 | Accuracy | 92.05% | 90.12% | 87.99% | 89.05% |
| 2 | Precision | 0.899 | 0.845 | 0.740 | 0.811 |
| 3 | Recall | 0.802 | 0.804 | 0.601 | 0.790 |

### 4.4 Comparative Analysis

The effect of hyperparameter tuning on the performances of the DL models is investigated by comparing the results of the performances of the optimized models with the results of the unoptimized models with respective to each of the three datasets considered in this study. That is, results in Table 6 are compared to the results in Table 9; results in Table 7 are compared to results in Table 10; and the results in Table 8 are compared with that of Table 11.

Tables 12, 13 and 14 present the corresponding differences in the performances of the models on the Twitter US airline dataset, stock market dataset, and IMDB movie review dataset respectively.

**Table 12: Performance Differences of the DL models on Twitter US Airline Dataset**

| S/N | Metrics | CNN-LSTM | CNN | LSTM | MLP |
|---|---|---|---|---|---|
| 1 | Accuracy | 1.07% | 2.01% | 0.96% | 1.04% |
| 2 | Precision | 0.099 | -0.005 | -0.01 | 0.021 |
| 3 | Recall | -0.06 | 0.005 | 0.019 | 0.102 |

**Table 13: Performance Differences of the DL models on Stock Market Dataset**

SLU JST
Sule Lamido University Journal of Science & Technology Vol. 7 No. 2 [December, 2023], pp. 139-159 https://doi.org/10.56471/
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

| S/N | Metrics | CNN-LSTM | CNN | LSTM | MLP |
|---|---|---|---|---|---|
| 1 | Accuracy | 1.92% | 4.39% | 1.31% | 1.66% |
| 2 | Precision | 0.016 | 0.021 | 0.01 | 0.012 |
| 3 | Recall | 0.01 | 0.025 | 0.009 | 0.014 |

**Table 14: Performance Differences of the DL models on IMDB Movie Review Dataset**

| S/N | Metrics | CNN-LSTM | CNN | LSTM | MLP |
|---|---|---|---|---|---|
| 1 | Accuracy | 1.86% | -0.89% | 1.00% | 3.07% |
| 2 | Precision | -0.091 | 0.005 | 0.06 | 0.091 |
| 3 | Recall | -0.008 | -0.006 | -0.109 | 0.1 |

Overall, the comparative results show that the difference in the performances of the optimized models versus the unoptimized model is marginal. The difference in the accuracy has a range of 5.28% with highest difference of 4.39% and lowest difference of -0.89%. The difference in precision has a range of 0.19 with highest difference of 0.099 and lowest difference of -0.091. Similarly, the difference in recall has a range of 20.211 with the highest difference of 0.102 and the lowest difference of -0.109.

Also, the results of the optimized models in this study are compared to the results in Mohamed et al. (2019) with respect to the IMDB dataset and the four DL models used in both studies. Table 15 presents the comparison between the results of the two studies. The table shows that all the optimized DL models in this study marginally outperform that of Mohamed et al. (2019) study. Given that the same dataset and DL models were used in both studies, the improvement in performance can be attributed to the hyperparameter tuning done using Randomsearchcv.

**Table 15: Comparison with Mohamed et al. (2019)**

| | CNN-LSTM | LSTM | MLP | CNN |
|---|---|---|---|---|
| **Mohamed et al. (2019)** | 89.20% | 86.64% | 86.74% | 87.7% |
| **Optimized DL models in this study** | 92.05% | 87.99% | 89.05% | 90.12% |
| **Marginal Improvement %** | **3.03%** | **1.35%** | **2.31%** | **2.42%** |

## 5. Conclusion

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Billions of textual data are being generated daily on social media, news publishing sites and ecommerce websites. Data mining has shown that valuable knowledge are hidden within data, and the use of mining techniques such as machine learning can uncover the hidden knowledge in data.

This study investigated the effect of hyperparameter optimization on the performances of some select DL models in opinion mining from three prominent textual datasets. CNN, LSTM, MLP and a hybridized CNN-LSTM models were studied using the random search technique for hyperparameter tuning. The three datasets used are Twitter US airline sentiment dataset, stock market dataset, and IMDB movie review dataset. Two hyperparameters, the number of epochs and batch size, were optimized for the DL models. Two sets of the DL models were implemented: the optimized DL models which have hyperparameter tunings and the unoptimized DL models which have no hyperparameter tuning. The effect of the hyperparameter tuning on the performance of the DL models was determined by comparing the performances of the optimized model to the performances of the unoptimized models. Also, this study compared the performances of the DL models with the results of the study reported in Mohamed et al. (2019) which also studied the same DL models not optimized and used the IMDB movie review dataset.

Overall, the results of the comparative analyses showed that the hyperparameter tuning marginally improved performance in majority of the cases within this study and when compared to the result reported in Mohamed et al. (2019). Also observed is that the hybridized model, CNN-LSTM, outperformed the other single models. Further studies can be conducted to investigate more ML or DL models both as single techniques and as hybridized techniques. Also, more tuning techniques can be applied to the models for hyperparameter optimization. Although, across the three datasets used in this study, the performances of the DL models were very similar, more datasets could be used in future studies to validate this similar performance observed in this study.

## References

Abo, M. E. M., Idris, N., Mahmud, R., Qazi, A., Hashem, I. A. T., Maitama, J. Z., Naseem, U., Khan, S. K., & Yang, S. (2021). A multi-criteria approach for arabic dialect sentiment analysis for online reviews: Exploiting optimal machine learning algorithm selection. In *Sustainability (Switzerland)* (Vol. 13, Issue 18). https://doi.org/10.3390/su131810018

Agarwal B., Mittal N. (2016) Machine Learning Approach for Sentiment Analysis. In: Prominent Feature Extraction for Sentiment Analysis. Socio-Affective Computing. Springer, Cham. https://doi.org/10.1007/978-3-319-25343-5_3

Ahmad, M., Aftab, S., Muhammad, S., & Ahmad, S. (2017). Machine Learning Techniques for Sentiment Analysis: A Review. Int. J. Multidiscip. Sci. Eng, 8(3), 27–32.

AlGhamdi, N., & Khatoon, S. (2021). Improving Sentiment Prediction using Heterogeneous and Homogeneous Ensemble Methods: A Comparative Study. Procedia Computer Science, 194, 60–68. https://doi.org/10.1016/j.procs.2021.10.059

Ali, K., Jamali, A., Abbas, M., Ali Memon, K., & Aleem Jamali, A. (2019). Multinomial Naive Bayes Classification Model for Sentiment Analysis. IJCS*NS International Journal of Computer Science and Network Security*, *19*(3), 62. https://doi.org/10.13140/RG.2.2.30021.40169

Ali Kandhro, I., Ameen Chhajro, M., Kumar, K., Lashari, H. N., & Khan, U. (2019). Student Feedback Sentiment Analysis Model Using Various Machine Learning Schemes A Review. *Indian Journal of Science and Technology*, *14*(12), 1–9. https://doi.org/10.17485/ijst/2019/v12i14/143243

Anuj, S., & Shubhamoy, D. (2012) (2012). *A comparative study of feature selection and machine learning techniques for sentiment analysis*. Proceedings of the 2012 ACM Research in Applied Computation Symposium. Retrieved January 25, 2022, from https://dl.acm.org/doi/abs/10.1145/2401603.2401605

Baid, P., Gupta, A., & Chaplot, N. (2017). Sentiment Analysis of Movie Reviews Using Machine Learning Techniques. *Lecture Notes in Networks and Systems*, *235*(December 2017), 361–369. https://doi.org/10.1007/978-981-16-2377-6_34

Bajeh, A.O., Alabidun, M.O., & Sadiku, P.O. (2018). Performance evaluation of some select machine learning algorithms for sentiment analysis. *International Journal of Information Processing and Communication, 6*(2), 409-419.

 

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Daudert, T. (2021). Exploiting textual and relationship information for fine-grained financial sentiment analysis. *Knowledge-Based Systems*, *230*, 107389. https://doi.org/10.1016/j.knosys.2021.107389

Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing*, *34*(6), 480–488. https://doi.org/10.1108/JCM-03-2017-2141

Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, *161*, 707–714. https://doi.org/10.1016/j.procs.2019.11.174

Erik., B., & Marie-Francine, M. (2008). A machine learning approach to sentiment analysis in multilingual Web text. *Information Retrieval*, *12*(5), 526–558. http://trec.nist.gov/pubs/trec15/t15

Fitri, V. A., Andreswari, R., & Hasibuan, M. A. (2019). Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm. *Procedia Computer Science*, *161*, 765–772. https://doi.org/10.1016/j.procs.2019.11.181

García-Ordás, M. T., Alaiz-Moretón, H., Benítez-Andrades, J. A., García-Rodríguez, I., García-Olalla, O., & Benavides, C. (2021). Sentiment analysis in non-fixed length audios using a Fully Convolutional Neural Network. *Biomedical Signal Processing and Control*, *69*. https://doi.org/10.1016/j.bspc.2021.102946

Garrick, R. C. (2020). *Enhanced Reader2.pdf*.

Gautam, G., & Yadav, D. (2014). Machine Learning Based Anxiety Prediction of General Public from Tweets During COVID-19. *Studies in Computational Intelligence*, *963*, 291–312. https://doi.org/10.1007/978-3-030-74761-9_13

Ghiassi, M., Zimbra, D., & Lee, S. (2016). Targeted Twitter Sentiment Analysis for Brands Using Supervised Feature Engineering and the Dynamic Architecture for Artificial Neural Networks. *Journal of Management Information Systems*, *33*(4), 1034–1058. https://doi.org/10.1080/07421222.2016.1267526

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine Learning-Based Sentiment Analysis for Twitter Accounts. *Mathematical and Computational Applications*, *23*(1), 11. https://doi.org/10.3390/mca23010011

IBM Cloud Education. (2020.). *What are neural networks?* IBM. Retrieved from: https://www.ibm.com/cloud/learn/neural-networks (Accessed on January 25, 2022)

Iqbal, F., Hashmi, J. M., Fung, B. C. M., Batool, R., Khattak, A. M., Aleem, S., & Hung, P. C. K. (2019). A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction. *IEEE Access*, *7*, 14637–14652. https://doi.org/10.1109/ACCESS.2019.2892852

Jain, A. P., & Dandannavar, P. (2016). Application of Machine Learning Techniques to Mineral Recognition. *Computer*, *October*, 628–632. http://innovexpo.itee.uq.edu.au/2001/projects/s369357/thesis.pdf

Kirilenko, A. P., Stepchenkova, S. O., Kim, H., & Li, X. (Robert). (2018). Automated Sentiment Analysis in Tourism: Comparison of Approaches. *Journal of Travel Research*, *57*(8), 1012–1025. https://doi.org/10.1177/0047287517729757

Krishna A., Akhilesh V., Aich A., Hegde C. (2019) Sentiment Analysis of Restaurant Reviews Using Machine Learning Techniques. In: Sridhar V., Padma M., Rao K. (eds) Emerging Research in Electronics, Computer Science and Technology. Lecture Notes in Electrical Engineering, vol 545. Springer, Singapore. https://doi.org/10.1007/978-981-13-5802-9_60.

Liu, C., Fang, F., Lin, X., Cai, T., Tan, X., Liu, J., & Lu, X. (2021). Improving sentiment analysis accuracy with emoji embedding. *Journal of Safety Science and Resilience*, *2*(4), 246–252. https://doi.org/10.1016/j.jnlssr.2021.10.003

Mariel, W. C. F., Mariyah, S., & Pramana, S. (2018). Sentiment analysis: A comparison of deep learning neural network algorithm with SVM and naïve Bayes for Indonesian text. *Journal of Physics: Conference Series*, *971*(1). https://doi.org/10.1088/1742-6596/971/1/012049

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mohamed Ali, N., El Hamid, M. M. A., & Youssif, A. (2019). Sentiment Analysis for Movies Reviews Dataset Using Deep Learning Models. *International Journal of Data Mining & Knowledge Management Process*, *09*(03), 19–27. https://doi.org/10.5121/ijdkp.2019.9302

Moraes, R., Valiati, J. F., & Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

comparison between SVM and ANN. *Expert Systems with Applications*, *40*(2), 621–633. https://doi.org/10.1016/j.eswa.2012.07.059

Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. *2013 4th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2013*. https://doi.org/10.1109/ICCCNT.2013.6726818

Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, *31*(1), 527–541. https://doi.org/10.1016/j.chb.2013.05.024

Phan, H. T., Tran, V. C., Nguyen, N. T., & Hwang, D. (2020). Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model. *IEEE Access*, *8*, 14630–14641. https://doi.org/10.1109/ACCESS.2019.2963702

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, *3*(2), 143–157. https://doi.org/10.1016/j.joi.2009.01.003

Selvi, C., Ahuja, C., & Sivasankar, E. (2015). A comparative study of feature selection and machine learning methods for sentiment classification on movie data set. *Advances in Intelligent Systems and Computing*, *343*, 367–379. https://doi.org/10.1007/978-81-322-2268-2_39

Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications*, *34*(4), 2622–2629. https://doi.org/10.1016/j.eswa.2007.05.028

Tran, D. D., Nguyen, T. T. S., & Dao, T. H. C. (2021). Sentiment Analysis of Movie Reviews Using Machine Learning Techniques. *Lecture Notes in Networks and Systems*, *235*(December 2021), 361–369. https://doi.org/10.1007/978-981-16-2377-6_34

Wu, J., Lu, K., Su, S., & Wang, S. (2019). Chinese Micro-Blog Sentiment Analysis Based on Multiple Sentiment Dictionaries and Semantic Rule Sets. *IEEE Access*, *7*, 183924–183939. https://doi.org/10.1109/ACCESS.2019.2960655

Zhao, Y., Dong, S., & Li, L. (2014). Sentiment analysis on news comments based on supervised learning method. *International Journal of Multimedia and Ubiquitous Engineering*, *9*(7), 333–346. https://doi.org/10.14257/ijmue.2014.9.7.28