# IJASA Article - Q3 - Lusiana - new - 041122

*by* Scientific Journal Reviewer

# Hybrid Modeling to Classify and Detect Outlier on Multilabel Dataset Based on Content and Context

Lusiana Efrizoni[1]
Faculty of Computer Science
STMIK Amik Riau
Indonesia
lusiana@stmik-amik-riau.ac.id

Sarjon Defit[2]
Faculty of Computer Science
UPI YPTK Padang
Indonesia
sarjon_defit@upiyptk.ac.id

Muhammad Tajuddin[3]
Faculty of Computer Science
Bumigora University
Indonesia
tajuddin@universitasbumigora.ac.id

***Abstract*— Text classification, almost all types of documents have been thoroughly studied in previous studies, but there is still little research for news categorization. It frequently suffers from ambiguity as a result of the different categories that fit and the poor dependability performance of most classification methods utilized, resulting in low efficiency. Most of the text classification tasks still use traditional Machine Learning algorithms. However, Machine Learning has weaknesses in training large-scale datasets and data sparseness often occurs from short texts. This study proposes a Hybrid model consisting of two models, namely the news classification model and the outlier detection model. Outlier detection is closely related to the text classification process. The news classification model using a combination of Long Short-Term Memory (LSTM)-Convolutional Neural Network (CNN, or ConvNet) and outlier classifier model is intended to predict which news is an outlier or not using a decision tree algorithm. The datasets used to train the model are AGNews and BBC News. The results of experiments conducted on two datasets: AGNews has an accuracy of 0.91 with a ROC Curve Score of 0.97 and BBC News has an accuracy value of 0.86 with an ROC Curve Score of 0.96. The LSTM-CNN hybrid model has excellent accuracy and AUC-ROC curve scores. As for the outlier classifier model, it can determine whether a news item is an outlier news or not with accuracy obtained 0.71 on BBC News and 0.92 on AGNews with the AUC-ROC curve score reaching 0.8 - 0.9.**

***Keywords—news article classification; hybrid model; outlier detection; ROC curve***

## I. INTRODUCTION

Digital news, often known as online news, is a type of modern news in which editorial information is provided over the Internet rather than through print or broadcast. Errors in grouping/classifying news might arise when using news or information from internet news sources. For example, news is categorized in the infotainment category, while based on the content of the news or the words contained in it, the news should be categorized in the politics category. Journalists and news monitoring companies (media monitoring companies) often face problems identifying topics in a very large number of news articles around the world [1]. Errors in categorizing or classifying information/news can also occur because the method used is still manually by reading the entire article to find the main topic. This method requires large resources and requires the reader's ability to extract the topic of a news/information document [2]. This fact shows that there is a discrepancy between the news category (or context) and the news content, or the meaning discussed or the news topic (as content) in categorizing or classifying news.

The increase in online news makes it difficult for internet users to access the content they are interested in, so it is necessary to classify news (text) so that it is easily accessible [3][4]. Coupled with the ever-growing volume of news corpus on the World Wide Web (WWW), it also creates problems in text classification, especially news article classification [5]. The categorization of news items frequently suffers from ambiguity as a consequence of the different categories that fit and the poor dependability performance of the majority of classification methods utilized, resulting in low efficiency [6]. Because human effort is no longer effective, automatic news (text) categorization must be created. People will not be prompted to think about which category the news belongs to if it is done automatically [2][7][8]. The capacity to categorize texts (news) into specific categories is extremely beneficial in dealing with information overload [9]. Text categorization is the process of assigning content-based labels to documents, which can be single-label (one text to one label) or multi-label (one text to several labels) [10]. Multilabel text classification is a task that involves categorizing text into one or more groups [11][12].

The Recurrent Neural Network (RNNrecurring )'s nature makes it well adapted for processing text with long variables, making it one of the most prominent architectures in Natural Language Processing (NLP) [11]. Meanwhile, the LSTM was created to overcome the gradient bursting and disappearing difficulties that often occur while training standard RNNs. Classification of online news texts using LSTM in this study because the LSTM structure is a full series or cannot be cut where the structure of the text document, if cut, would affect the meaning of the phrase. Before categorizing the text, the usage of word embedding will be an input feature in the LSTM. Aside from RNN/LSTM, CNN has shown remarkable achievements in the field of natural language processing. To get strong results in topic classification and semantic analysis, CNN can be integrated with word vectors [13]. The CNN input matrix retrieves just the word vector matrix from the word detail level and ignores the overall semantic feature expression from the text breakdown level, resulting in an inadequate representation of text features and potentially affecting text classification accuracy [14].

This study will look at how an extracted summary of the text (news content) and context (similarity of words and connections between words) might assist readers or consumption models filter vital information from the text when classifying online news in English. The word embedding method is used to represent words in vector space in content-based and context-based representations. The methods used in content-based representation are Latent Semantic Analysis (LSA) and Singular Value Decomposition (SVD), while for context-based representation, the technique used is Word2Vec. In addition to the classification performance of news articles, outlier detection is also the focus of this research. Because outlier detection is very closely related to the text classification process [15]. Outliers are abnormal patterns or events that do not match the expected events or patterns [16]. Outlier detection is used to detect news that does not fit the category of news articles. The findings of the study are meant to be utilized by news management organizations on online news portals to filter outlier news items (deviations/falsehoods) that are discovered. Furthermore, automated news/information categorization is crucial for dealing with multi-label news articles on internet sites [9].

## II. Related Works

Previous scholars have conducted text classification study in accordance with the method of classifying texts or online news, such as:

- Stein et al [17] used the RCV1 dataset to assess presentation text (i.e. GloVe, word2vec, and fastTex) using a mixture of classification models (i.e. fastText, XGBoost, SVM, and CNN) for hierarchical text categorization (HTC). According to the study results, FastText is a classification approach that produces extremely good results as word embedding, despite the fact that the amount of data presented is quite minimal. Precision, recall, and F1 values are 0.920, 0.922, and 0.920, respectively.
- Researchers [18] tested the word2vec and doc2vec features on a clinical text classification task and compared the results to the classic bag-of-words (BOW) feature. Learning reveals that the word2vec feature outperforms the BOW-1-gram feature. In combination sets bigger than six modalities (Acupuncture, Biofeedback, Guided Imagery, Meditation, Tai-Chi, and Yoga), BOW-1,2-gram performs better, with an AUC value of 0.91 and an accuracy of 0.85. Imagery that is guided. Meanwhile, in the smaller individual modality set, word2vec outperforms, with AUC values ranging from 0.80 to 0.93 and accuracy ranging from 0.82-0.86. The information utilized by Veterans Affairs (VA) electronic medical records (EMR) is housed in the Veterans Administration Informatics and Computing Infrastructure (VINCI) database.
- The study includes the LSTM2 model for document categorization, which includes repLSTM for adaptive data representation and rankLSTM for integrated learning ranking. In repLSTM, the supervised LSTM is used to examine document representation by introducing label documents. When the semantics are coherent with and conform to LSTM sequential learning, the semantic tree is utilized to reorder the order of document labels in rankLSTM. The word embedding is BoW, and the dataset comes from Bio (10C), email, and News. In document classification tasks, the model produces F1 Measure outcomes of less than 75% [19].
- The researcher suggests combining the word2vec neural network model and the Latent Dirichlet Allocation (LDA) document topic model to create a text feature. The matrix model is represented by Word2Vec and LDA. Text classification studies were carried out after the feature matrix was fed into a convolutional neural network (CNN) for convolution pooling. The Sogou Corpus Text Classification Lab provided the experimental data, which included 8,000 papers from sports, military, tourism, finance, IT, real estate, education, and entertainment, as well as eight categories of 1000 experiments. The experimental findings reveal that the suggested matrix model outperforms the standard text classification methods based on word2vec and CNN. The recall rate and F1 of the three evaluation indicators rose by 8.4%, 8.9%, and 8.4%, respectively, at the level of text classification accuracy [14].
- The author proposes a weighted word2vec model for emotion classification that integrates an Attention mechanism into the Long-Short Term Memory (LSTM) model. The weight matrix is utilized with TFIDF to produce the LSTM input after word2vec encodes the text information into the word vector. English and Chinese data were used in the analysis. The English dataset consists of 25,000 movie data points from IMDB, each comprising 12,500 positive and negative text values. The Chinese dataset, which comprises 6000 words and 3,000 positive and negative evaluations, was constructed from hotel review corpora (Chn Senti Corp.). According to the testing data, this approach has an accuracy, recall, and F1 measure of 0.87 [20].
- Following study [19] looked on label embedding for text representation and suggested a label-embedding attentive model. The approach embeds words and labels in the same merged space and assesses the compatibility of word-label pairings in order to address document representation. The learning system was evaluated on 5 datasets and clinical text apps (AGNews, Yelp Review Full, Yelp Review Polarity, DBPedia, and Yahoo! Answers Topic). The investigation's findings demonstrate that the suggested LEAM (Label-Embedding Attentive Model) approach has much reduced computing costs and outperforms CNN, LSTM, Simple Word Embeddings-based Models (SWEM), and bi-directional blocks. network of self-attention (Bi-BloSAN). Predictive performance is as follows [21], with F1 and micro mean areas of 0.91 and 0.88 under the ROC (AUC) curve, and precision at n (P@n) of 0.61.
- By taking into account the combined weighting factors of grammatical categories and high frequency topic terms, the researchers developed a short text categorization approach based on word vectors and the suggested LDA topic model. Gibbs sampling is employed in this approach to train the LDA topic model depending on the weight of the segment of

speech. The model is tested with the word vector Wor2vec and vectorized with high-frequency themes. The ex-tend text function is then tested. The SVM algorithm was used to classify the expanded short text after increasing the feature, and the classification results were assessed using Precision (83.6), F1-score (84.4), and recall (84.4). (85.4). The dataset is generated from Sogou Lab's news corpus, which has 6,000 titles extracted and classified into six categories: computer, health, sports, tourism, education, and military. Each field comprises 1000 articles of fewer than 200 words that are text data brief [22].

- Researchers propose a novel topic-based skip-gram neural language model to investigate topic-based word embedding for indexing biomedical literature with CNN. Topic-based skip-grams combine textual material with topic models such as Latent Dirichlet Allocation (LDA) to capture topic-based exact word connections and then incorporate them into distributed word embedding learning. The combination of topic-based Skip-grams with multimodal CNN architecture outperforms advanced approaches in indexing biomedical literature, annotating clinical records, and general textual dataset categorization. The F1 score of 82.7% was used to assess the model's performance [23].
- This work introduces a novel active learning strategy for text classification. The primary objective of active learning is to decrease labeling effort while maintaining classification accuracy by intelligently selecting which samples to label. The recommended technique chooses an informative sample set based on the posterior probability provided by a sequence of multi-class SVM classifiers, which is subsequently manually labeled by an expert. The datasets used are from the text category (TC), namely the Reuters-21578 document (R8), the 20ng dataset, and the WebKB collection. TFIDF implements word embedding. The accuracy of each dataset varies: R8 (83.33%), 20ng (43.79%), and WebKB (53.77%) [24].
- The study examined four well-known news categorization algorithms: Nave Bayes, SVM, Random Forest, and MLP Classifier. Because of its homogeneity, Nave Bayes is likely to be a superior way to serve as a text categorization model than other approaches. This research provides a news classification comparing four classifiers in which numerous different forms of news such as business and finance, sports, politics and policy, criminal justice, and health have been categorized [3].
- The researchers assessed the performance of the categorization system using the Scopus dataset. The two challenges that affect performance in different methods in text classification are classification and feature extraction from documents utilizing the derived features. The performance of classification algorithms such as Nave Bayes (NB) and K-Nearest Neighbor (K-NN) improves when Bayesian boost and bagging are used. To increase the performance of the text classification algorithm, data preparation and cleaning methods are applied to the selected data set, and class imbalance concerns are investigated. Nave Bayes has an overall accuracy of 71.11%, whereas KNN has an accuracy of 78.67%. The experimental findings demonstrate that KNN outperforms NB. [25].
- Researchers investigated the best method for automatically classifying news items in Indonesian. The dataset was obtained from www.cnnindonesia.com using web crawling. The document is first subjected to several Text Preprocessing methods (such as Lemmatization and Stopwords Removal), followed by feature selection algorithms (such as the TF-IDF and SVD algorithms) and classification algorithms for Multinomial Nave Bayes, Multivariate Bernoulli Nave Bayes, and Support Vector Machine. According to the findings of the tests, the combination of TF-IDF and Multinomial Nave Bayes Classifier produces the best results when compared to other algorithms, with a precision of 0.9841519 and a recall of 0.9840000. Previous comparable study that classified Indonesian-language news stories with an accuracy of 85% was surpassed by the findings. [2].
- The researcher assesses the accuracy and efficiency of Kolmogorov Complexity Distance Measure (KCDM) and Artificial Neural Network (ANN) computing time for large-dimensional news article categorization issues. British Broadcasting Corporation (BBC) News utilized this dataset. The dataset contains 2225 news stories divided into five categories: politics (417), sports (511), entertainment (386), education and technology (401), and business (401). (510). After tokenization and stop-word deletion, Porter's technique was used for word stemming, and Normalized Term Frequency-Inverse Document Frequency (NTF-IDF) was employed for feature extraction. The experimental findings reveal that ANN outperforms KCDM in terms of accuracy while KCDM outperforms ANN in terms of computational time efficiency [5].
- Researchers [26] address the challenges of the text categorization process. The experiment was done out with the RNN+LSTM+GRU model implemented. This model is pitted against RCNN+LSTM and RNN+GRU. The GloVe dataset was used to test the models. The accuracy and recall of the models were evaluated. The F1 score was used to compare the two models' performance. The hybrid RNN model has three LSTM layers and two GRU layers, whereas the RCNN model includes four convolution layers and four LSTM levels and the RNN model includes four GRU layers. The weighted average for the hybrid RNN model was 0.74, 0.69 for RCNN+LSTM, and 0.77 for RNN+GRU. The RNN+LSTM+GRU model has poor accuracy in the first epoch but gradually improves with and as the epoch advances.
- The authors [26] offer a Hybrid RNN-based Deep Learning Approach for Text Classification in which the Traditional RNN Model is combined with the RNN+LSTM+GRU model. The RNN+LSTM+GRU model was built and compared to the Region-Based

Convolutional Neural Network (RCNN) + LSTM and RNN+GRU models. The GloVe dataset was used to train the models. The accuracy and recall parameters are used to assess the model's performance. The F1 Score is used to compare the performance of the model. The hybrid RNN model consists of three LSTM layers and two GRU layers, whereas the RCNN model consists of four convolution layers and four LSTM levels and the RNN model consists of four GRU layers. The weighted average for the hybrid RNN model was 0.74, 0.69 for the RCNN+LSTM model, and 0.77 for the RNN+GRU model. The RNN+LSTM+GRU model has moderate accuracy at the start of the epoch, but it gradually improves as the epoch length rises.

From several related studies (review papers) on text classification, machine learning algorithms are generally used for text classification combined with word embedding models. The use of a deep learning classification approach that combines RNN/LSTM and CNN for text classification is currently restricted. In this paper, we present a hybrid modeling (RNN/LSTM + CNN) with word embedding feature (i.e. Word2vec + LSA/SVD) for news text classification, along with a model for outlier detection utilizing machine learning methods, particularly the Decision Tree approach.

## III. METHODOLOGY

The hybrid model developed consists of 2 models, namely a classification model using a deep learning algorithm (i.e. RNN/LSTM+CNN) and an outlier detection model using a machine learning algorithm (i.e. Decision Tree). LSTM-CNN is used to classify news articles based on Word2Vec-based context and LSA/SVD-based content. The outlier detection model is intended to predict which news is an outlier or not. Before the dataset is ready to be used for modeling, the dataset needs to be processed in several stages, such as text processing, data sharing into training and testing data, and feature extraction (word embedding) contained in news texts. When the data is ready for use, it is used for model training and data testing to confirm the model's forecast findings. Fig. 1 depicts the hybrid concept in broad strokes..

Labeling whether a news item is an outlier or not, the classification results from the LSTM+CNN model are used as input in the outlier detection process. The classification results generated by the model, there will be errors (Outliers). The use of the term topic in research, what is meant is content. For example, if a news item has topic A, but the model predicts the probability that the news will go to topic B. Hybrid LSTM-CNN model is used to classify the context and content of a news, while the outlier classifier model is intended to predict which news is an outlier or not. Before the data can be utilized for modeling, it must go through various phases of processing, including text processing, splitting the data into training and testing data, and extracting the characteristics included in the news content. When the data is ready for use, it is used for model training and data testing to check the model's prediction findings. There will be a 0.6 mistake if subject A is at number 0.4. The root mean square error (RMSE) for the full dataset may be calculated using the error computation. Outliers are samples or raw data with a larger error than the RMSE. Outlier labeling that has been done, allows to form an outlier classification model.

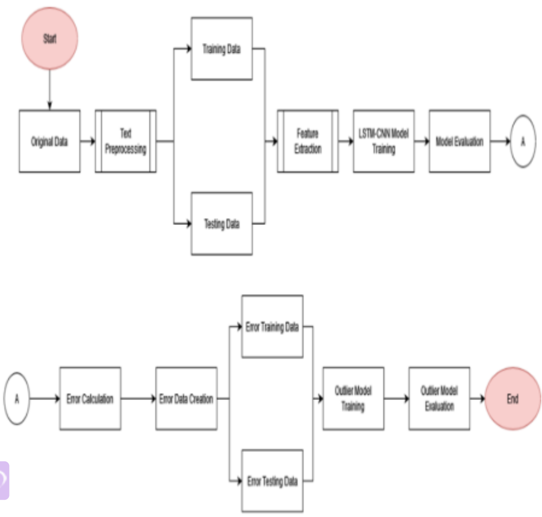


Fig. 1. Hybrid Model Framework

### A. Dataset

This study's text dataset is a public dataset (AGNews (or ODDs) and BBC News). a) According to Zhang and LeCun (2015), AGNews is divided into four categories: World (152,469), Entertainment (115,967), Sports (108,344), and Business (45,639). (Wang et al, 2018). Over 400,000 text data samples are included in the AGNews dataset. b) Greene and Cunningham's British Broadcasting Corporation (BBC) news dataset (2006). The dataset contains 2225 news items divided into five categories: politics (417), sports (511), entertainment (386), education and technology (401), and business (401). (510). Articles were collected from the BBC news website between 2004 and 2005 and made freely available at http://mlg.ucd.ie/datasets/bbc.htm.

### B. Text Preprocessing

The phases of text preparing news articles from raw data to ready-to-use data are as follows: elimination of symbols and numbers, tokenization, deletion of stop words, and lemmatization (as shown in Fig. 2). The elimination of symbols and numerals is done since they do not have a specific significance that is relevant to the news issue. Tokenization is then used to break down phrases or paragraphs in the news into pieces of words that the model can read. Furthermore, to decrease noise in the data, meaningless terms such as subject (I, he, she, etc.) must be deleted. Finally, lemmatization is employed to shorten word forms like eating to eat.

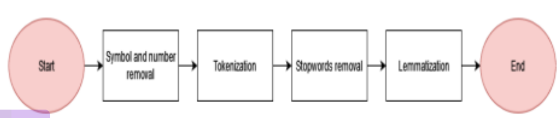


Fig. 2. Stages of Text preprocessing

### C. *Feature extraction*

The model is still unable to fully utilize the cleaned news text. Feature extraction is a method for extracting information from text. Fig. 3 depicts the steps of feature extraction. This study employs two extraction strategies, namely context and content-based information/news extraction. The Word2Vec model was utilized to extract context-based information. Each word from the preceding text processing is turned into a 100-dimensional vector. To obtain the vector, a window of value 5 is employed, which indicates that 5 surrounding words will be used to interpret the context of a word in the paragraph. The Word2Vec model produces vectors with dimensions of 100, with the number of vectors equaling the number of words in the training data.

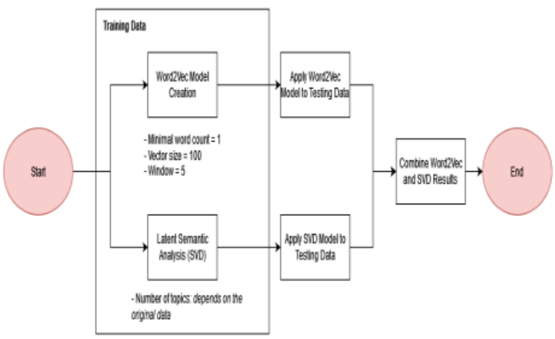


Fig. 3. Stages of Feature Extraction

The singular value decomposition (SVD) approach is used to extract content-based news. This approach attempts to divide the word matrix generated from text processing results into three matrices: a matrix carrying the relationship between news and words, a matrix providing the relationship between news and topics, and information representing the link between topics and words. The connection matrix between themes and words is employed. As a result, each word has a vector with information about its link to themes. After obtaining two types of vectors from the Word2Vec and SVD findings, the two outputs are concatenated into one vector. This vector is known as the embedding vector, and it represents one word.

### D. *Hybrid model of LSTM+CNN architecture*

Figure 3.7 presents the architecture of the classification model of the Hybrid model using 2 (two) deep learning algorithms (RNN/LSTM+CNN).

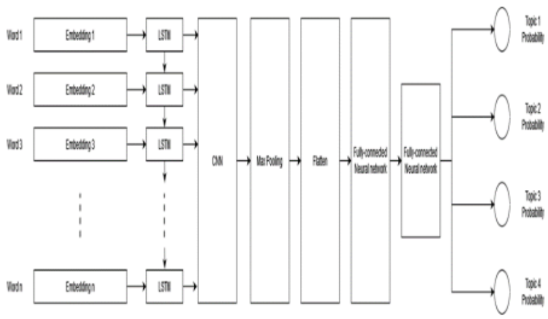


Fig. 4. Hybrid Model Architecture

The feature extraction embedding vector is utilized as an embedding layer in LSTM-CNN modeling. When a news story is utilized as model input, each word in the story is turned into the matching embedding vector. The embedding vector is then used as the LSTM input layer. The LSTM cells utilized were 64 in number. With a filter size of 100, a kernel size of 2, and a ReLu activation function, the output of each LSTM cell will be utilized as the CNN input layer. To decrease noise in the output, the CNN output matrix must be introduced into the Max Pooling layer. After reducing the noise in the output, the output, which has been turned into a 1-dimensional vector in the flatten layer, may be fed into a fully connected neural network (NN). In this case, we utilize two NN layers, the first with 16 nodes and the ReLu activation function, and the second with the number of nodes according to the number of themes in the data. In the second layer, the Softmax activation function is employed to generate output in the form of the probability of each topic (content).

### E. *Model Evaluation*

The receiver operating characteristic (ROC) curve is a graphical depiction of the classifier's performance. The ROC curve is created by computing and displaying the True Positive Rate (TPR) vs. False Positive Rate (FPR) for a single classifier at various thresholds (Threshold). Formulae 1 and 2 show the TPR and FBR equations.

$$TPR=Sensitivity=TP/(TP+TN) \quad (1)$$

$$FPR=1\text{-}specificity=FP/(FP+TN) \quad (2)$$

The number of True Positives is TP, whereas the number of False Negatives is FN. TPR is a measure of the likelihood that a true positive event will be categorized as such. While FP represents the number of False Positives, TN represents the number of True Negatives. The FPR is a measure of how frequently a "false alarm" occurs, or how frequently a real negative event is categorized as positive. AUC (area under the (ROC) curve) is used to visualize the ROC curve. The better the classifier performs for a particular task, the higher the AUC value. Figure 5 depicts the ROC Curve with AUC Score. Table shows the categorization category based on the ROC Curve score [27].


Fig. 5. ROC Curve with AUC Score

In general, an AUC value of 0.5 indicates poor classification performance (failure), 0.7 - 0.8 is regarded excellent, 0.8 - 0.9 is considered very good, and more than 0.9 is considered outstanding. Table 1 displays the Model Category Based on ROC Curve Value [28].

TABLE I. MODEL CATEGORY BASED ON ROC CURVE VALUE

| Accuracy | Category |
|---|---|
| 0.90 – 1.00 | Excellent Classification |
| 0.80 – 0.90 | Good Classification |
| 0.70 – 0.80 | Fair Classification |
| 0.60 – 0.70 | Poor Classification |
| 0.50 – 0.60 | Failure Classification |

### *F. Outlier Model Evaluation*

This step employs the LSTM-CNN model prediction results to determine if a news item is an outlier or not. The model's prediction outputs will contain mistakes that can be computed. For example, if a news item includes subject A but the model forecasts that the likelihood of the news entering topic A is just 0.4, the error is 0.6. We can determine the root mean square error (RMSE) of the complete dataset from this computation, and data with an error greater than RMSE will be classified as outliers, as indicated in the table above. The outlier labeling helps us to create an outlier classifier model.

## IV. RESULT AND DISCUSSION

### *A. News Classification*

The architecture used to build the model with the AGNews dataset includes: (1) an embedding matrix with an input size of 100 (the result of the previous stage's extraction of content and context information); (2) an LSTM with 64 nodes; and (3) a CNN with the following parameters: filter = 100, kernel = 2, number of strides = 1, Padding = valid, activation function = Relu, (4) Flattening; (5) Pooling; (6) Dense hidden layer with 16 nodes and relu activation function; (7) Dropout with rate = 0.5; and (8) Dense hidden layer with 4 nodes (depending on goal number) and softmax activation function. Meanwhile, the hybrid architectural model for the BBC News dataset, points 1–7, is identical to the AGNews dataset, with the exception of point 8. BBC News employs a dense hidden layer with a node count of 5. (according to the number of targets). This step involves doing hyperparameter tweaking to obtain the optimal collection of hyperparameters.

The model must now be fitted. Model training is carried out at this level. If there is no rise in the accuracy value in testing data testing after 10 iterations (epochs), the training will be terminated, and the modeling results will be saved as shown in Table 2.

TABLE 2. HYPERPARAMETERS

| Layer (type) | Output Shape | Param # |
|---|---|---|
| embedding_2 (Embedding) | (None, 100, 104) | 6533800 |
| lstm_2 (LSTM) | (None, 100, 96) | 77184 |
| conv1d_2 (Conv1D) | (None, 97, 100) | 38500 |
| global_max_pooling1d_2 (GlobalMaxPooling1D) | (None, 100) | 0 |
| flatten_2 (Flatten) | (None, 100) | 0 |
| dense_4 (Dense) | (None, 10) | 1010 |
| dense_5 (Dense) | (None, 4) | 44 |

Total params: 6,650,538
Trainable params: 6,650,538
Non-trainable params: 0

Following the construction of the hybrid model using training data and hyperparameter adjustment. The Confusion Matrix (CM) and ROC Curve were used to evaluate the model (AUC). After 235 rounds, the test results reveal a loss value of 0.2818, an accuracy of 0.9065, a test loss of 0.28183448, and a test acc of 0.90653336. As demonstrated in Table 3, the CM on the Model using the AGNews dataset reveals that the data was accurately predicted (TP) for each actual label.

*TABLE 3. CM OF AGNEWS DATASET*

| Accuracy 90,65 | | | | | |
|---|---|---|---|---|---|
| ***Confusion Matrix*** | **Label** | ***World*** | ***Entertainment*** | ***Sport*** | ***Business*** |
| | ***World*** | 6620 | 232 | 373 | 256 |
| | ***Entertainment*** | 65 | 7250 | 45 | 73 |
| | ***Sport*** | 199 | 83 | 6703 | 595 |
| | ***Business*** | 219 | 62 | 602 | 6623 |

The evaluation model using the Confusion Matrix on BBC News is presented in Table 4.

*TABLE 4. CM OF BBC NEWS DATASET*

| Accuracy 0.85 | | | | | | |
|---|---|---|---|---|---|---|
| ***Confusion Matrix*** | **Label** | ***Entertainment*** | ***Tech*** | ***Politics*** | ***Business*** | ***Sport*** |
| | ***Entertainment*** | 64 | 2 | 1 | 14 | 0 |
| | ***Tech*** | 2 | 52 | 0 | 6 | 0 |
| | ***Politics*** | 0 | 2 | 49 | 12 | 0 |
| | ***Business*** | 3 | 2 | 5 | 76 | 0 |
| | ***Sport*** | 1 | 0 | 1 | 5 | 76 |

The accuracy of the test on the AGNews dataset is 91%, the ROC Curve (AUC) value is 0.9832, and the training validation has a loss value of 0.1146 and an accuracy of 0.9611 (96.11%). The accuracy in training, testing, and validation results, as well as the ROC curve value, suggest that the model performs effectively. The ROC curve for AGNews is shown in Fig. 6.

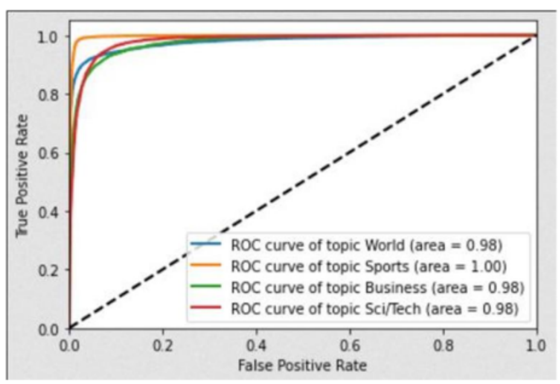


Fig. 6. ROC Curve of the AGNews Dataset

The ROC Curve of the BBC News dataset is presented in Fig. 7.

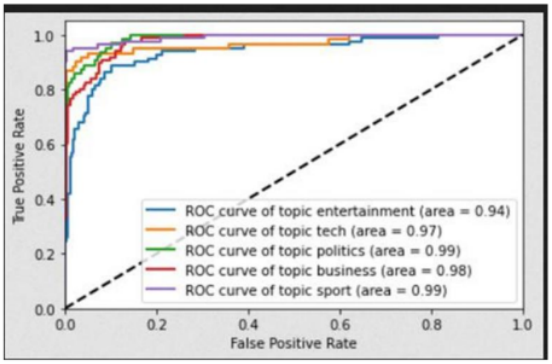


Fig. 7. ROC Curve from BBC News Dataset

The ROC curve depicts the relationship between the true positive rate (TPR) and the false positive rate (FPR). The classifier with the curve closest to the top left corner (perfect classifier) performs best. The closer the curve is to the ROC space's 45-degree diagonal, the less accurate the classifier.

TABLE 5. PERFORMANCE EVALUATION OF TWO DATASET

| Model | Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|
| AG News | World | 0.93 | 0.88 | 0.91 | 7481 |
| | Entertainment | 0.95 | 0.98 | 0.96 | 7433 |
| | Sport | 0.87 | 0.88 | 0.88 | 7580 |
| | Business | 0.88 | 0.88 | 0.88 | 7506 |
| | | | Accuracy | 0.91 | 30000 |
| **Model** | **Class** | **Precision** | **Recall** | **F1-score** | **Support** |
| BBC News | entertainment | 0.91 | 0.79 | 0.85 | 81 |
| | Tech | 0.90 | 0.87 | 0.88 | 60 |
| | Politics | 0.80 | 0.78 | 0.82 | 6 |
| | Business | 0.67 | 0.88 | 0.76 | 86 |
| | Sport | 1.00 | 0.92 | 0.96 | 83 |
| | | | Accuracy | 0.85 | 373 |

The model is trained and tested (tested) against two datasets: AGNews and BBC News. The model assessment results demonstrate that the LSTM+CNN hybrid model has high accuracy and ROC AUC ratings (see Table 5).

## B. Outlier Detection

The initial step in detecting outliers is to analyze faulty data. At this point, the error model's prediction outputs will be gathered into 1 (one) independent data frame comprised of test subjects and train topics. The amount of the error is estimated using the equation: error=1-max(xi), i1,2,3,4, where x i is the greatest probability score assigned by the model. Each prediction error in this formulation may be determined by the size of the error, allowing the RMSE error data to be calculated. The error size of the RMSE dataset is 0.93. Fig. On the AGNews dataset, Fig. 8 is error data train and Fig. 9 is error data test.

| | Text | Actual Label | Topic 0 score | Topic 1 score | Topic 2 score | Topic 3 score | Topic | Error | outlier_label |
|---|---|---|---|---|---|---|---|---|---|
| 2182 | [president, hugo, chavez, rid, high, overwhelm... | 0 | 0.03 | 0.00 | 0.96 | 0.00 | 2 | 0.04 | Normal |
| 75249 | [caetano, defender, serginho, take, hospital, ... | 0 | 0.15 | 0.85 | 0.00 | 0.00 | 1 | 0.15 | Normal |
| 57388 | [european, union, agree, lift, long, stand, sa... | 0 | 0.24 | 0.00 | 0.75 | 0.00 | 2 | 0.25 | Normal |
| 38321 | [brussels, european, commission, confirm, want... | 3 | 0.02 | 0.00 | 0.72 | 0.26 | 2 | 0.28 | Normal |
| 49457 | [initial, public, offer, share, newly, establi... | 3 | 0.04 | 0.00 | 0.59 | 0.37 | 2 | 0.41 | Outlier |

Fig. 8. Error_data_train on AGNews

| | Text | Actual Label | Topic 0 score | Topic 1 score | Topic 2 score | Topic 3 score | Topic | Error | outlier_label |
|---|---|---|---|---|---|---|---|---|---|
| 82611 | [personal, england, wales, another, record, hi... | 2 | 0.12 | 0.61 | 0.23 | 0.05 | 1 | 0.39 | Outlier |
| 8679 | [bangladesh, captain, bashar, rule, next, mont... | 0 | 0.40 | 0.59 | 0.00 | 0.00 | 1 | 0.41 | Outlier |
| 44571 | [internet, phone, service, sipphone, charge, v... | 2 | 0.00 | 0.00 | 0.09 | 0.90 | 3 | 0.10 | Normal |
| 81887 | [people, least, hear, still, think, intend, sp... | 2 | 0.10 | 0.02 | 0.31 | 0.57 | 3 | 0.43 | Outlier |
| 98948 | [cincinnati, still, among, dangerous, cities, ... | 2 | 0.28 | 0.01 | 0.32 | 0.40 | 3 | 0.60 | Outlier |

Fig. 9. Error_data_test on AGNews

The amount of data that has an error greater than the RMSE is around 65%. Data that has an error of more than RMSE will be considered as an outlier, and others as normal. Figure 10 is error_data_test, on the AGNews dataset.

| | Text | Actual Label | Topic 0 score | Topic 1 score | Topic 2 score | Topic 3 score | Topic 4 score | Topic | Error |
|---|---|---|---|---|---|---|---|---|---|
| 1422 | [firm, embrace, ecommerce, firm, embrace, inte... | 2 | 0.00 | 0.99 | 0.00 | 0.01 | 0.00 | 1 | 0.01 |
| 1071 | [allow, scrutiny, urge, give, watchdogs, freed... | 0 | 0.00 | 0.98 | 0.02 | 0.00 | 0.00 | 1 | 0.02 |
| 450 | [national, gallery, pink, national, gallery, h... | 0 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1 | 0.00 |
| 264 | [franz, seek, government, help, franz, ferdina... | 0 | 0.12 | 0.88 | 0.00 | 0.00 | 0.00 | 1 | 0.12 |
| 1072 | [game, maker, fight, survival, britain, larges... | 1 | 0.00 | 0.25 | 0.00 | 0.75 | 0.00 | 3 | 0.25 |
| 186 | [telegraph, newspapers, job, daily, sunday, te... | 3 | 0.00 | 0.96 | 0.00 | 0.04 | 0.00 | 1 | 0.04 |
| 1087 | [blue, slam, blackburn, savage, birmingham, co... | 4 | 0.00 | 0.59 | 0.15 | 0.00 | 0.26 | 1 | 0.41 |
| 89 | [campbell, lions, consultant, former, governme... | 4 | 0.00 | 0.00 | 0.79 | 0.00 | 0.20 | 2 | 0.21 |
| 664 | [orange, colour, clash, court, colour, orange,... | 3 | 0.00 | 0.98 | 0.01 | 0.00 | 0.00 | 1 | 0.02 |
| 819 | [rule, tackle, weddings, rule, marriage, forei... | 2 | 0.00 | 0.97 | 0.01 | 0.02 | 0.00 | 1 | 0.03 |
| 1337 | [glastonbury, fan, card, fan, ticket, year, gl... | 0 | 0.00 | 0.61 | 0.31 | 0.08 | 0.00 | 1 | 0.39 |
| 136 | [portfolio, sale, collection, murder, fashion,... | 0 | 0.04 | 0.93 | 0.00 | 0.00 | 0.03 | 1 | 0.07 |
| 369 | [lead, interactive, bafta, win, national, thea... | 1 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0 | 0.00 |

Fig. 10. Error_data_test on BBC News

Two sorts of methods are created to forecast if a news item is an outlier or not. The outlier classifier model is deployed independently from the LSTM-CNN hybrid model in the first system (Fig.11). The meta-modeling approach is applied in the second system (Fig. 12), specifically the outlier classifier model, which uses the prediction outputs of another model (the LSTM-CNN hybrid model in this case) to identify whether a news item is an outlier or not. In both types of systems, decision tree models are utilized to assess which system performs the best at detecting outliers.

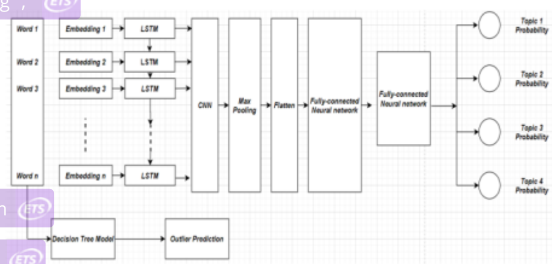


Fig. 11. The first model of the classification of news outliers

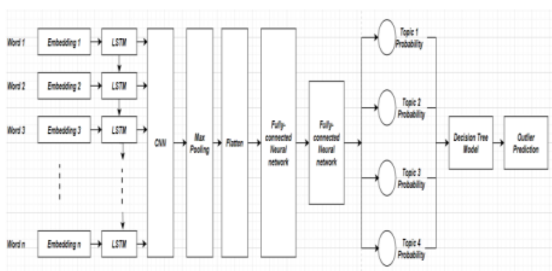


Fig. 12. The second model of the classification of news outliers

### C. Validation model

Fig. 13 depicts the validation models for the two categorization models. It is clear that system 2 outperforms system 1 for both types of datasets. The ROC AUC value in System 1 is only about 0.5-0.6, indicating that the model predictions are not significantly different from the random prediction outcomes. However, in system 2, the ROC AUC score may reach 0.8 - 0.9, indicating that by using the LSTM-CNN hybrid model's probability prediction findings, the outlier classifier model can correctly assess whether a news item is an outlier or not.

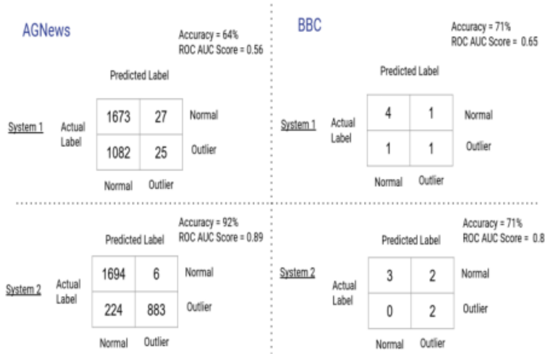


Fig. 13. Comparison of Outlier news classifications from two datasets

### D. Comparative Study

The findings of the hybrid model's performance evaluation include two models: the LSTM-CNN hybrid model and the outlier classifier model. To classify based on context (i.e. feature extraction Word2Vec) and content (i.e. feature extraction SVD), the hybrid LSTM-CNN model is utilized. The combination of background and content is referred to as a subject in the context of news. The outlier categorization approach is designed to predict whether or not a piece of news is an outlier.

As a classification of multilabel text from the Word2Vec model, the hybrid model is 100-dimensional vectors, with the number of vectors equal to the number of words in the training data. A singular value decomposition (SVD) model is used to extract content-based news. After obtaining two types of vectors from the Word2Vec and SVD findings, the two outputs are concatenated into one vector. This vector is known as the embedding vector, and it represents one word.

Table 6 compares the current research to earlier studies. The performance evaluation of the proposed hybrid model's training and testing yielded improved results. The hybrid model-built falls into the Good Classification group, according to the ROC Curve. This demonstrates that the Hybrid model with RNN/LSTM+CNN architecture and feature extraction Wor2Vec+LSA/SVD can be utilized to categorize sequential text effectively. Furthermore, the created Hybrid model is capable of detecting news outliers.

TABLE 6. COMPARISON WITH PREVIOUS RESEARCH

| Word Embedding | Text Classification Strategies | Dataset | Result |
|---|---|---|---|
| GloVe [26] | • RNN + LSTM + GRU<br>• RCNN+ LSTM<br>• RNN+GRU | GloVe | Accuracy 0.74 Accuracy 0.69 Accuracy 0.77 |
| GloVe, word2vec, and fastText [17] | CNN, SVM, XGBoost | RCV1 | Precision value 0.92, Recall 0.92 and F1 measure 0.920 |
| Word2Vec, doc2vec, BoW [18] | SVM | VINCI | AUC value between 0.80 -0.93 and accuracy value between 0.82-0.86 |
| BoW [19] | LSTM2: repLSTM and rankLSTM | Bio (10C), email and News | F1 Measure less than 75% |
| Word2Vec +LDA [14] | CNN | Sogou Corpus text classification Lab | Accuracy 0.84, recall 0.89 and f1-score 0.86 |
| Word2Vec + TF-IDF [20] | Att-LSTM | IMDB film review and hotel review corpora | Precision value 0.87, recall 0.87 and F1 measure 0.87 |
| [29] | Label-Embedding Attentive Model (LEAM) | AGNews, Yelp Review Full, Yelp Review Polarity, DBPedia and Yahoo! Answers Topic | Under the ROC (AUC) curve, the F1 micro average is 0.91 and the macro average is 0.88, with a precision of 0.61. |
| LDA + Word2Vec [22] | SVM | Corpus News | Precision value (83.6), F1-score (84.4), and recall (85.4). |
| LDA [23] | Skip-Gram and CNN | Biomedical literature, clinical record annotation | F1 score 82.7%. |
| TF-IDF [24] | SVM | Reuters-21578 document (R8), 20ng dataset and WebKB collection | The accuracy of the dataset varies R8 (83.33%), 20ng (43.79%), and WebKB (53.07 |
| [3] | Naïve Bayes, SVM, Random Forest, MLP Classifier | News categories | The Support Vector Classifier has the highest accuracy of 0.6134 |
| TF-IDF [25] | Naïve Bayes (NB) and K-Nearest Neighbor (K-NN) | Corpus Dataset | Naïve Bayes accuracy is 71.11%, and KNN 78.67%. |
| TF-IDF dan SVD [2] | Multinomial Nave Bayes, Multivariate Bernoulli Naïve Bayes, and Support Vector Machine | CNN indonesia | When compared to other algorithms, TF-IDF and Multinomial Nave Bayes Classifier produce the best results, with accuracy of 0.9841519 |

| | | | |
|---|---|---|---|
| | | | and recall of 0.9840000. |
| Normalized TF-IDF (NTF-IDF) [5] | Kolmogorov Complexity Distance Measure (KCDM) and ANN | British Broadcasting Corporation (BBC) News | In terms of accuracy, ANN outperforms KCDM, whereas KCDM outperforms ANN in terms of computational time efficiency. |
| **Word2Vec+LSA/SVD** | **RNN/LSTM+CNN** | **AGNews BBC News** | **AGNews Accuracy 0.91; BBC News accuracy 0.85; AGNews ROC Curve between 0.98 – 1.00; BBC News ROC Curve between 0.94 – 0.99.** |
| **Word2Vec+LSA/SVD** | **News Classification (RNN/LSTM+CNN) & Outlier Detection (architecture 2)** | **AGNews BBC News** | **AGNews: Accuracy 0.92 and ROC Curve 0.89 BBC News: 0.71 Accuracy and ROC Curve 0.8** |

The last two are the results of the research conducted (bold).

## V. CONCLUSION

The created hybrid approach is made up of two models: the news categorization model (News Categories) and the outlier classification model. Based on the context and content of a news item, the news categorization model employs a deep learning algorithm (i.e. LSTM+CNN). While the outlier classifier model is designed to predict whether or not a piece of news is an outlier. AGNews and BBC News datasets were utilized for modeling. The Word2Vec model was used to extract context-based news, while the Singular Value Decomposition (SVD) model was utilized to extract content-based news. AGNews' trials yielded an accuracy of 0.91 with a ROC Curve Score of 0.97, whereas BBC News yielded an accuracy of 0.86 with a ROC Curve Score of 0.96. The hybrid model offers high accuracy and ROC AUC values.

In the process of categorizing a news item as having outliers or not, the projected data from the LSTM-CNN model is used and evaluated using two types of models. First, the outlier classifier model was deployed independently from the LSTM-CNN hybrid model. The second model employs the meta-modeling approach, with the outlier classifier model utilizing the prediction findings of another model (LSTM-CNN hybrid model). The decision tree method is utilized in the outlier classification model. The second model examined is considerably superior to the previous one. The ROC AUC score in the first model is only about 0.5-0.6, indicating that the model predictions are not significantly different from the random prediction outcomes. However, in the second model, the ROC AUC score may reach 0.8 - 0.9, indicating that by using the LSTM-CNN hybrid model's probability prediction outputs, the outlier classifier model can accurately assess whether a news item is an outlier or not. scored 0.71 points on BBC News and 0.92 points on AGNews.

## REFERENCES

[1] D. Liparas, Y. HaCohen-Kerner, A. Moumtzidou, S. Vrochidis, dan I. Kompatsiaris, "News articles classification using random forests and weighted multimodal features," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8849, no. October, hal. 63–75, 2014, doi: 10.1007/978-3-319-12979-2_6.

[2] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, dan Rudy, "News Article Text Classification in Indonesian Language," *Procedia Comput. Sci.*, vol. 116, hal. 137–143, 2017, doi: 10.1016/j.procs.2017.10.039.

[3] P. K. Verma, "Exploration of text classification approach to classify news classification," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 5, hal. 2555–2562, 2020.

[4] A. M. Gaber, A. M. Gaber, dan H. Moussa, "SMAD: Text Classification of Arabic Social Media Dataset for News Sources," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 10, hal. 508–516, 2021, doi: 10.14569/IJACSA.2021.0121058.

[5] T. Matthew Fagbola, C. Surendra Thakur, dan O. Olugbara, "NEWS ARTICLE CLASSIFICATION USING KOLMOGOROV COMPLEXITY DISTANCE MEASURE AND ARTIFICIAL NEURAL NETWORK," *Int. J. Technol.*, vol. 10, no. 4, hal. 710–720, 2019.

[6] S. Biradar dan M. M. Raikar, "Performance Analysis of Text Classifiers based on News Articles-A Survey," *Indian J. Sci. Res.*, vol. 15, no. 2, hal. 156–161, 2017.

[7] A. Z. Woldaregay *dkk*, "Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes," *Artif. Intell. Med.*, vol. 98, hal. 109–134, 2019, doi: https://doi.org/10.1016/j.artmed.2019.07.007.

[8] A. Abdullahi, N. A. Samsudin, S. K. A. Khalid, dan Z. A. Othman, "An Improved Multi-label Classifier Chain Method for Automated Text Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 3, hal. 442–449, 2021, doi: 10.14569/IJACSA.2021.0120352.

[9] I. C. Irsan dan M. L. Khodra, "Hierarchical multi-label news article classification with distributed semantic model based features," *Int. J. Adv. Intell. Informatics*, vol. 5, no. 1, hal. 40–47, 2019, doi: 10.26555/ijain.v5i1.168.

[10] R. Romero, P. Celard, J. M. Sorribes-Fdez, A. Seara Vieira, E. L. Iglesias, dan L. Borrajo, "MobyDeep: A lightweight CNN architecture to configure models for text classification," *Knowledge-Based Syst.*, vol. 257, hal. 109914, 2022, doi: 10.1016/j.knosys.2022.109914.

[11] W. K. Sari, D. P. Rini, dan R. F. Malik, "Text Classification Using Long Short-Term Memory With GloVe Features," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 5, no. 2, hal. 85, 2020, doi: 10.26555/jiteki.v5i2.15021.

[12] W. Liu, J. Pang, N. Li, X. Zhou, dan F. Yue, "Research on Multi-label Text Classification Method Based on tALBERT-CNN," *Int. J. Comput. Intell. Syst.*, vol. 14, no. 1, 2021, doi: 10.1007/s44196-021-00055-4.

[13] K. Chen, Z. Zhang, J. Long, dan H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Syst. Appl.*, vol. 66, hal. 1339–1351, 2016, doi: 10.1016/j.eswa.2016.09.009.

[14] M. Gao, T. Li, dan P. Huang, *Text classification research based on improved word2vec and CNN*, vol. 11434 LNCS. Springer International Publishing, 2019.

[15] A. Noori, S. S. B. Kamaruddin, dan F. B. K. Ahmad, "Towards an outlier detection model in text data stream," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 6, hal. 2970–2973, 2019, doi: 10.30534/ijatcse/2019/47862019.

[16] P. Krammer, O. Habala, J. Mojžiš, L. Hluchý, dan M. Jurkovič, "Anomaly Detection Method for Online Discussion," *Procedia Comput. Sci.*, vol. 155, hal. 311–318, 2019, doi: https://doi.org/10.1016/j.procs.2019.08.045.

[17] A. J. Stein, J. Weerasinghe, S. Mancoridis, dan R. Greenstadt, "News Article Text Classification and Summary for Authors and Topics," hal. 1–12, 2020, doi: 10.5121/csit.2020.101401.

[18] Y. Shao, S. Taylor, N. Marshall, C. Morioka, dan Q. Zeng-Treitler, “Clinical Text Classification with Word Embedding Features vs. Bag-of-Words Features,” *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018*, hal. 2874–2878, 2019, doi: 10.1109/BigData.2018.8622345.

[19] Y. Yan, Y. Wang, W. C. Gao, B. W. Zhang, C. Yang, dan X. C. Yin, “LSTM 2: Multi-Label Ranking for Document Classification,” *Neural Process. Lett.*, vol. 47, no. 1, hal. 117–138, 2018, doi: 10.1007/s11063-017-9636-0.

[20] H. Yuan, Y. Wang, X. Feng, dan S. Sun, “Sentiment analysis based on weighted word2vec and ATT-LSTM,” *ACM Int. Conf. Proceeding Ser.*, hal. 420–424, 2018, doi: 10.1145/3297156.3297228.

[21] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, dan J. Eisenstein, “Explainable prediction of medical codes from clinical text,” *NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, hal. 1101–1111, 2018, doi: 10.18653/v1/n18-1100.

[22] F. Sun dan H. Chen, “Feature extension for Chinese short text classification based on LDA and Word2vec,” *Proc. 13th IEEE Conf. Ind. Electron. Appl. ICIEA 2018*, no. 1, hal. 1189–1194, 2018, doi: 10.1109/ICIEA.2018.8397890.

[23] H. Xu, A. Kotov, M. Dong, A. I. Carcone, D. Zhu, dan S. Naar-King, “Text classification with topic-based word embedding and Convolutional Neural Networks,” *ACM-BCB 2016 - 7th ACM Conf. Bioinformatics, Comput. Biol. Heal. Informatics*, no. April 2019, hal. 88–97, 2016, doi: 10.1145/2975167.2975176.

[24] M. Goudjil, M. Koudil, M. Bedda, dan N. Ghoggali, “A Novel Active Learning Method Using SVM for Text Classification,” *Int. J. Autom. Comput.*, vol. 15, no. 3, hal. 290–298, 2018, doi: 10.1007/s11633-015-0912-z.

[25] M. Azam, T. Ahmed, F. Sabah, dan M. I. Hussain, “Feature Extraction based Text Classification using K-Nearest Neighbor Algorithm,” *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 18, no. 12, hal. 95–101, 2018, [Daring]. Tersedia pada: http://paper.ijcsns.org/07_book/201812/20181213.pdf.

[26] P. Sunagar dan A. Kanavalli, “A Hybrid RNN based Deep Learning Approach for Text Classification,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 6, hal. 289–295, 2022.

[27] J. N. Mandrekar, “Receiver operating characteristic curve in diagnostic test assessment,” *J. Thorac. Oncol.*, vol. 5, no. 9, hal. 1315–1316, 2010, doi: 10.1097/JTO.0b013e3181ec173d.

[28] F. Gorunescu, *Data Mining: Concepts, models and techniques. Vol (12)*. Springer Science & Business Media, 2011.

[29] G. Wang *dkk*, “Joint embedding of words and labels for text classification,” *ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 1, hal. 2321–2331, 2018, doi: 10.18653/v1/p18-1216.

# IJASA Article - Q3 - Lusiana - new - 041122

ORIGINALITY REPORT

| 15% | 11% | 7% | 4% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to Hoa Sen University<br>Student Paper | 2% |
| 2 | link.springer.com<br>Internet Source | 1% |
| 3 | thesai.org<br>Internet Source | 1% |
| 4 | ijtech.eng.ui.ac.id<br>Internet Source | 1% |
| 5 | Haotian Xu, Ming Dong, Dongxiao Zhu, Alexander Kotov, April Idalski Carcone, Sylvie Naar-King. "Text Classification with Topic-based Word Embedding and Convolutional Neural Networks", Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2016<br>Publication | 1% |
| 6 | Fanke Sun, Heping Chen. "Feature extension for Chinese short text classification based on LDA and Word2vec", 2018 13th IEEE | <1% |

Conference on Industrial Electronics and Applications (ICIEA), 2018
Publication

| | | |
|---|---|---|
| 7 | www.warse.org<br>Internet Source | <1% |
| 8 | Submitted to Nepal College of Information Technology<br>Student Paper | <1% |
| 9 | journal.uad.ac.id<br>Internet Source | <1% |
| 10 | Yijun Shao, Stephanie Taylor, Nell Marshall, Craig Morioka, Qing Zeng-Treitler. "Clinical Text Classification with Word Embedding Features vs. Bag-of-Words Features", 2018 IEEE International Conference on Big Data (Big Data), 2018<br>Publication | <1% |
| 11 | coderzcolumn.com<br>Internet Source | <1% |
| 12 | mdpi-res.com<br>Internet Source | <1% |
| 13 | search.ijcsns.org<br>Internet Source | <1% |
| 14 | Submitted to Bilkent University<br>Student Paper | <1% |

| | | |
|---|---|---|
| 15 | Mengyuan Gao, Tinghui Li, Peifang Huang. "Chapter 11 Text Classification Research Based on Improved Word2vec and CNN", Springer Science and Business Media LLC, 2019<br>Publication | <1 % |
| 16 | journal.universitasbumigora.ac.id<br>Internet Source | <1 % |
| 17 | sjia.ejournal.unsri.ac.id<br>Internet Source | <1 % |
| 18 | Submitted to University of Westminster<br>Student Paper | <1 % |
| 19 | ieeexplore.ieee.org<br>Internet Source | <1 % |
| 20 | Submitted to Michigan Technological University<br>Student Paper | <1 % |
| 21 | peerj.com<br>Internet Source | <1 % |
| 22 | www.tandfonline.com<br>Internet Source | <1 % |
| 23 | doaj.org<br>Internet Source | <1 % |
| 24 | www.ineteconomics.org<br>Internet Source | <1 % |

| | | |
|---|---|---|
| 25 | "ECAI 2020", IOS Press, 2020<br>Publication | <1 % |
| 26 | Francesco Gargiulo, Stefano Silvestri, Mario Ciampi, Giuseppe De Pietro. "Deep neural network for hierarchical extreme multi-label text classification", Applied Soft Computing, 2019<br>Publication | <1 % |
| 27 | R. Romero, P. Celard, J.M. Sorribes-Fdez, A. Seara Vieira, E.L. Iglesias, L. Borrajo. "MobyDeep: A lightweight CNN architecture to configure models for text classification", Knowledge-Based Systems, 2022<br>Publication | <1 % |
| 28 | arxiv.org<br>Internet Source | <1 % |
| 29 | cora.ucc.ie<br>Internet Source | <1 % |
| 30 | oaktrust.library.tamu.edu<br>Internet Source | <1 % |
| 31 | stoptb.org<br>Internet Source | <1 % |
| 32 | www.ijcseonline.org<br>Internet Source | <1 % |
| 33 | www.ijraset.com<br>Internet Source | <1 % |

34 Foulkes, Andrea S., Gregory J. Matthews, Ujjwal Das, Jane F. Ferguson, Rongheng Lin, and Muredach P. Reilly. "Mixed Modeling of Meta-Analysis P-Values (MixMAP) Suggests Multiple Novel Gene Loci for Low Density Lipoprotein Cholesterol", PLoS ONE, 2013.
Publication
<1 %

---

35 jurnal.politeknik-kebumen.ac.id
Internet Source
<1 %

---

36 www.aimspress.com
Internet Source
<1 %

---

37 www.hindawi.com
Internet Source
<1 %

---

38 Ilyos Rabbimov, Sami Kobilov, Iosif Mporas. "Uzbek News Categorization using Word Embeddings and Convolutional Neural Networks", 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT), 2020
Publication
<1 %

---

39 Minni Jain, Harshita Rastogi. "Automatic Text Summarization using Soft-Cosine Similarity and Centrality Measures", 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020
Publication
<1 %

---

| | | |
|---|---|---|
| 40 | ijiset.com<br>Internet Source | <1 % |
| 41 | muhammadrizkifblog.wordpress.com<br>Internet Source | <1 % |
| 42 | shunk031.me<br>Internet Source | <1 % |
| 43 | www.mdpi.com<br>Internet Source | <1 % |
| 44 | bth.diva-portal.org<br>Internet Source | <1 % |
| 45 | dokumen.pub<br>Internet Source | <1 % |
| 46 | ejurnal.stmik-budidarma.ac.id<br>Internet Source | <1 % |
| 47 | isyou.info<br>Internet Source | <1 % |
| 48 | norma.ncirl.ie<br>Internet Source | <1 % |
| 49 | www.intechopen.com<br>Internet Source | <1 % |
| 50 | www.researchgate.net<br>Internet Source | <1 % |
| 51 | Eun Hak Lee, Seung-Young Kho, Dong-Kyu Kim, Shin-Hyung Cho. "Travel time Prediction | <1 % |

Using Gated Recurrent Unit and Spatio-temporal Algorithm", Proceedings of the Institution of Civil Engineers - Municipal Engineer, 2020
Publication

---

| | | |
|---|---|---|
| 52 | "A Modified Screening Tool for Autism (Checklist for Autism in Toddlers [CHAT-23]) for Chinese Children", PEDIATRICS, 08/01/2004<br>Publication | <1 % |
| 53 | Thomas Adams, Stefano Giani, William M. Coombs. "A high-order elliptic PDE based level set reinitialisation method using a discontinuous Galerkin discretisation", Journal of Computational Physics, 2019<br>Publication | <1 % |

---

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | Off |
| Exclude bibliography | On | | |

# IJASA Article - Q3 - Lusiana - new - 041122

PAGE 1

**S/V** This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

**Article Error** You may need to use an article before this word. Consider using the article **the**.

**Article Error** You may need to use an article before this word. Consider using the article **the**.

**Confused** You have used either an imprecise word or an incorrect word.

**Article Error** You may need to remove this article.

**Wrong Form** You may have used the wrong form of this word.

**Article Error** You may need to use an article before this word.

**Missing ","** Review the rules for using punctuation marks.

**Article Error** You may need to use an article before this word.

PAGE 2

**S/V** This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

**Article Error** You may need to use an article before this word. Consider using the article **the**.

**P/V** You have used the passive voice in this sentence. You may want to revise it using the active voice.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**P/V** You have used the passive voice in this sentence. You may want to revise it using the active voice.

**Article Error** You may need to remove this article.

**Missing ","** Review the rules for using punctuation marks.

**Article Error** You may need to use an article before this word.

**Article Error** You may need to use an article before this word.

**Article Error** You may need to remove this article.

**Missing ","** Review the rules for using punctuation marks.

**Article Error** You may need to use an article before this word.

**Article Error** You may need to remove this article.

**Article Error** You may need to use an article before this word. Consider using the article **the**.

**Missing ","** Review the rules for using punctuation marks.

**Article Error** You may need to use an article before this word.

**Frag.** This sentence may be a fragment or may have incorrect punctuation. Proofread the sentence to be sure that it has correct punctuation and that it has an independent clause with a complete subject and predicate.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Article Error** You may need to use an article before this word.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**P/V** You have used the passive voice in this sentence. You may want to revise it using the active voice.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**S/V** This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

**Missing ","** Review the rules for using punctuation marks.

**Missing ","** Review the rules for using punctuation marks.

**Article Error** You may need to use an article before this word.

**P/V** You have used the passive voice in this sentence. You may want to revise it using the active voice.

**Article Error** You may need to use an article before this word.

**Article Error** You may need to use an article before this word.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**P/V** You have used the passive voice in this sentence. You may want to revise it using the active voice.

**P/V** You have used the passive voice in this sentence. You may want to revise it using the active voice.

**Article Error** You may need to use an article before this word.

**Missing ","** Review the rules for using punctuation marks.

**Prep.** You may be using the wrong preposition.

**Article Error** You may need to remove this article.



**S/V** This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

**Article Error** You may need to use an article before this word.

ETS Missing "," Review the rules for using punctuation marks.

ETS Missing "," Review the rules for using punctuation marks.

ETS Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

ETS Sentence Cap. Review the rules for capitalization.

ETS Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

ETS Confused You have used either an imprecise word or an incorrect word.

ETS S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.



ETS S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

ETS Article Error You may need to use an article before this word.

ETS Article Error You may need to remove this article.

ETS Word Error Did you type **the** instead of **they**, or have you left out a word?

ETS Article Error You may need to use an article before this word.

ETS P/V You have used the passive voice in this sentence. You may want to revise it using the active voice.

ETS Article Error You may need to use an article before this word.

ETS Article Error You may need to use an article before this word.

ETS Wrong Form You may have used the wrong form of this word.

**S/V** This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

**Wrong Form** You may have used the wrong form of this word.

**Missing ","** Review the rules for using punctuation marks.

**Confused** You have used either an imprecise word or an incorrect word.

**S/V** This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

**S/V** This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**S/V** This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Article Error** You may need to remove this article.

**Article Error** You may need to use an article before this word.

**Article Error** You may need to use an article before this word.

**Article Error** You may need to use an article before this word. Consider using the article **a**.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Article Error** You may need to use an article before this word.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Frag.** This sentence may be a fragment or may have incorrect punctuation. Proofread the sentence to be sure that it has correct punctuation and that it has an independent clause with a complete subject and predicate.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Missing ","** Review the rules for using punctuation marks.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Missing ","** Review the rules for using punctuation marks.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**S/V** This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.



**S/V** This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

**Article Error** You may need to use an article before this word.

**Missing ","** Review the rules for using punctuation marks.

**Possessive**

**Article Error** You may need to use an article before this word.

**Missing ","** Review the rules for using punctuation marks.

**Wrong Form** You may have used the wrong form of this word.

**Missing ","** Review the rules for using punctuation marks.

**Article Error** You may need to use an article before this word. Consider using the article **the**.

**Wrong Form** You may have used the wrong form of this word.

**Article Error** You may need to use an article before this word. Consider using the article **a**.

**Article Error** You may need to remove this article.

**Proper Nouns** You may need to use a capital letter for this proper noun.

**P/V** You have used the passive voice in this sentence. You may want to revise it using the active voice.

**Article Error** You may need to use an article before this word.

**Article Error** You may need to use an article before this word. Consider using the article **the**.



---

**S/V** This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

**Article Error** You may need to remove this article.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Wrong Article** You may have used the wrong article or pronoun. Proofread the sentence to make sure that the article or pronoun agrees with the word it describes.

**Article Error** You may need to use an article before this word. Consider using the article **the**.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Wrong Article** You may have used the wrong article or pronoun. Proofread the sentence to make sure that the article or pronoun agrees with the word it describes.

**Article Error** You may need to use an article before this word.

**Missing ","** Review the rules for using punctuation marks.

**Sp.** This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

**Wrong Article** You may have used the wrong article or pronoun. Proofread the sentence to make sure that the article or pronoun agrees with the word it describes.

**Article Error** You may need to use an article before this word. Consider using the article **the**.

**Wrong Form** You may have used the wrong form of this word.

**Proper Nouns** You may need to use a capital letter for this proper noun.

**Proper Nouns** You may need to use a capital letter for this proper noun.



---

**S/V** This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

**Wrong Form** You may have used the wrong form of this word.

**Article Error** You may need to remove this article.

**Article Error** You may need to use an article before this word. Consider using the article **the**.

**Article Error** You may need to use an article before this word. Consider using the article **the**.