Hybrid Modeling to Classify and Detect Outliers on Multilabel Dataset based on Content and Context

By Muhammad Tajuddin

WORD COUNT 7421
TIME SUBMITTED 03-JUN-2023 09:44AM
PAPER ID 100200789

Hybrid Modeling to Classify and Detect Outliers on Multilabel Dataset based on Content and Context

Lusiana Efrizoni[1], Sarjon Defit[2], Muhammad Tajuddin[3]
Faculty of Computer Science, STMIK Amik Riau, Pekanbaru, Indonesia[1]
Faculty of Computer Science, UPI YPTK Padang, Padang, Indonesia[2]
Faculty of Computer Science, Bumigora University, Nusa Tenggara Barat, Indonesia[3]

***Abstract*—Due to the linked various matching categories, news article categorization are a rapidly increasing field of interest in text classification. However, the low-reliability indices and ambiguities related to frequently used province classifiers restrict success in this field. Most of the existing research uses traditional machine learning algorithms. It has weaknesses in training large-scale datasets, and data sparseness often occurs from short texts. Therefore, this study proposed a hybrid model consisting of two models, namely the news article classification and the outlier detection model. The news article classification model used a combination of two deep learning algorithms (Long Short-Term Memory dan Convolutional Neural Network) and outlier classifier model, which was intended to predict the outlier news using a decision tree algorithm. The proposed model's performance was compared against two widely used datasets. The experimental results provide useful insights that open the way for a number of future initiatives.**

Keywords*—News article classification; machine learning; outlier detection

I. Introduction

Digital or online news is a form of contemporary news where editorial content is distributed over the internet as opposed to published via print or broadcast. News or information contained in online news portals allows errors to occur in grouping/classifying news. 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 manual 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 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]. 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 [4]. Readers can receive much information on the online news portal. Sometimes, they take it for granted without any selection of information. On that basis, much of the current information media classifies the categorization before dissemination to the general public. This classification is useful to make it easier for people to find the desired information. The classification of news articles often suffers from ambiguity due to the various categories that fit and the weak reliability performance of most classification systems used, resulting in low efficiency [5]. Therefore, automatic news (text) classification needs to be developed because manual work is no longer effective. If it is done automatically, people will not be asked to think about which category the news belongs to [2]. The ability to classify texts (news) into certain categories is very helpful in dealing with information overload [6]. Multilabel text classification is an activity to categorize text into one or more categories [7].

Recurrent Neural Network (RNN) is one of the most popular architectures used in Natural Language Processing (NLP) because the recurring structure is suitable for processing text with long variables [7]. Meanwhile, the Long Short-Term Memory (LSTM) was developed to solve the exploding and vanishing gradient problems that can be encountered during traditional RNNs training. Classification of online news texts using LSTM was utilized in this study because the LSTM structure is a complete series or cannot be cut since cutting text document structure changes the meaning of the sentence. Word embedding was used as an input feature in the LSTM before classifying the text. Apart from RNN/LSTM, Convolutional Neural Network (CNN) has also achieved excellent results in the NLP field. CNN can be combined with word vectors in topic classification and semantic analysis to achieve good results [8]. The CNN input matrix only extracts the word vector matrix from the word detail level and ignores the overall semantic feature expression from the text breakdown level, which leads to an incomplete representation of text features and can affect the accuracy of text classification [9].

This study aims to evaluate how extractive summary of the text (news content) and context (similarity of words and relationships between words) can help filter important information from the text either for readers or consumption models in 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 method used in content-based representation is Latent

Semantic Analysis (LSA) and Singular Value Decomposition (SVD), while in 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 since outlier detection is very closely related to the text classification process [10]. Outliers are abnormal patterns or events that do not match the expected events or patterns [11]. Outlier detection is used to detect news that does not fit the category of news articles. The research results are expected to be used by news management agencies on online news portals to classify news articles according to their categories and filter or block if they are found to contain outliers. In addition, automatic news/information categorization is very important for handling multi-label news article classification in online portals [6].

The remainder of this work is structured systematically. The second section provides brief overviews of text classification (or news article categorization) approaches. The research approach for the experiments is presented in Section III. The outcomes of the experiments are reported in Section IV, while the conclusion is presented in Section V.

II. Related Works

Relevant research related to text classification has been carried out in previous studies in accordance with the process of classifying texts or online news articles, including:

- The research of Stein et al. [12] analyzed presentation text (i.e., GloVe, word2vec, and fastTex) with a combination of classification models (i.e., fastText, XGBoost, SVM and CNN) for hierarchical text classification (HTC) using the RCV1 dataset. The analysis results showed that fastText was a classification method that provided very good results as word embedding, although the amount of data provided was relatively small. The precision, recall and F1 measure values were 0.920, 0.922, and 0.920, respectively. In contrast, our study focuses on deep learning to classify news articles based on their categories and machine learning for outlier detection.
- Shao et al. [13] experimented with the word2vec and doc2vec features for a clinical text classification task and compared the results with the traditional bag-of-words (BOW) feature. Learning showed that the word2vec feature performed better than the BOW-1-gram feature. In combination sets that were larger than six modalities (i.e., acupuncture, biofeedback, guided imagery, meditation, tai-chi and yoga), BOW-1,2-gram had better performance compared to the other feature extraction, with an area under curve (AUC) value of 0.91 and an accuracy of 0.85 specifically for the guided imagery. Meanwhile, in the smaller individual modality set, word2vec performed better compared to the other feature extraction, with AUC values between 0.80-0.93 and accuracy values between 0.82-0.86. The dataset used by Veterans Affairs (VA) electronic medical records (EMR) was stored in the Veterans Administration Informatics and Computing Infrastructure (VINCI) database. In contrast, this research only utilized word2vec in feature extraction.
- Research by Yan et al. [14] proposed the LSTM2 model for document classification consisting of repLSTM for the adaptive data representation process and rankLSTM for the integrated learning ranking process. In repLSTM, the supervised LSTM was used to study document representation by inserting label documents. In rankLSTM, the order of document labels was rearranged according to the semantic tree, where the semantics were compatible with and conformed to LSTM sequential learning. The word embedding used was BoW, and the dataset was taken from Bio (10C), email and News. The model achieved F1 Measure results of less than 75% in document classification tasks [14]. Meanwhile, the research in the article builds the news article classification model using LSTM+CNN.
- Gao et al. [9] proposed a text feature that combined the word2vec neural network model and the Latent Dirichlet Allocation (LDA) document topic model. Word2Vec and LDA represented the matrix model. The feature matrix was entered into a CNN for convolution pooling, and text classification experiments were performed. The experimental data come from the Sogou corpus text classification lab with 8,000 documents from sports, military, tourism, finance, IT, real estate, education, entertainment, and eight categories of 1000 experiments. The experimental results suggested that the proposed matrix model had a better classification effect than the traditional text classification method based on word2vec and CNN. At the level of text classification accuracy, the recall rate and F1 of the three evaluation indicators increased by 8.4%, 8.9% and 8.6% [9]. The similarity between the previous and current studies is that both use Word2Vec and LDA in feature extraction.
- Other research conducted by Yuan et al. [15] proposed a weighted word2vec, adding an attention mechanism to the LSTM model for emotion classification. After the text information was encoded into the word vector by word2vec, the weight matrix was combined with TFIDF to form the LSTM input. The dataset used is English data set and Chinese data set. The English dataset is an IMDB film review set consisting of 25,000 movie data, positive and negative values of 12,500 texts each. The Chinese dataset was collected from the hotel review corpora (Chn Senti Corp.) with 6000 data where the positive and negative values were 3,000 texts each. The experimental results showed that this method had a precision, recall and F1 measure of 0.87, respectively [15]. In contrast, the research in the article builds the news article classification model using LSTM+CNN.
- Subsequent research by Wang et al. [14] investigated label embedding for text representation and proposed a label-embedding attentive model. The model

embedded words and labels in the same merged space and measured the compatibility of word-label pairs to attend to document representation. The learning framework was tested on five datasets (i.e., AGNews, Yelp Review Full, Yelp Review Polarity, DBPedia and Yahoo! Answers Topic) and clinical text applications. The investigation results indicated that the proposed LEAM (Label-Embedding Attentive Model) algorithm required much lower computational costs, and achieved better performance compared to CNN, LSTM, Simple Word Embeddings-based Models (SWEM) and bi-directional blocks self-attention network (Bi-BloSAN). Following predictive performance [16], F1 and micro mean area was 0.91 and macro average was 0.88 under the ROC (AUC) curve, and precision at n (P@n) was 0.61. In contrast, the
research in the article builds the news article classification model using LSTM+CNN. The similarity between the previous and current studies is the use of ROC (AUC) curve to evaluate the model.

- Sun & Chen [17] designed a short text classification method based on word vectors and the proposed LDA topic model by considering the combined weighting factors of grammatical categories and high frequency topic words. In this method, Gibbs sampling was used to train the LDA topic model based on the weight of the part of speech. The model was exercised using the word vector Wor2vec and vectorized with high frequency topics. Then, the ex-tend text feature was tested. After expanding the feature, the SVM algorithm was used to classify the extended short text, and the classification results were evaluated using Precision (83.6), F1-score (84.4), and recall (85.4). The dataset was taken from the news corpus provided by Sogou Lab, which consisted of a total of 6,000 titles after being extracted, divided into six categories: computer, health, sports, tourism, education and military. Each category had 1000 essays, which was less than 200 words and belonged to text data short [17]. The similarity between the previous and current studies is that both used Word2Vec and LDA in feature extraction, but the algorithm used in classification was different.

- Xu et al. [18] proposed a new topic-based skip-gram neural language model to study topic-based word embedding for indexing biomedical literature with CNN. Topic-based skip-grams utilized textual content with topic models, for example, LDA, to capture topic-based precise word relationships and then integrated them into distributed word embedding learning. The combination of topic-based Skip-grams and multimodal CNN architecture outperformed advanced methods in indexing biomedical literature, annotating clinical records, and general textual dataset classification. The performance of the model was measured using the F1 score with a value of 82.7% [18]. In contrast, the research in the article builds the news article classification model using LSTM+CNN.

- In this study, we present a new active learning method for text categorization. The main goal of active learning is to reduce labeling effort without compromising classification accuracy by intelligently choosing which samples to label. The proposed method selects an informative sample set using the posterior probabilities provided by a set of multi-class SVM classifiers, and these samples are then manually labeled by an expert. The datasets used are public datasets in the text category (TC), namely Reuters-21578 document (R8), 20ng dataset and WebKB collection. Word embedding used was TFIDF. The accuracy of each dataset varies where R8, 20ng and WebKB had values of 83.33%, 43.79% and 53.07%, respectively [19].

- Verma [3] compared four very prominent algorithms for news classification, namely Naïve Bayes (NB), SVM, Random Forest and Multi-layer Perceptron (MLP) Classifier. Compared to the other approaches, Naïve Bayes is likely to be a better approach to serve as a text classification model because of its homogeneity. The paper proposed news classification by comparing four classifiers in which several different types of news have been classified, such as business & finance, sports, politics & policy, criminal justice, and health [3]. In contrast, the research in the article builds the news article classification model using LSTM+CNN.

- Next, Azan et al. [20] analyzed the performance of the classification algorithm using the Scopus dataset. In text classification, classification and feature extraction from documents using the extracted features were the main problems in reducing performance of different algorithms. The performance of classification algorithms such as NB and K-Nearest Neighbor (K-NN) showed a better improvement using Bayesian boost and bagging. Data preprocessing and cleaning steps were induced on the selected data set, and class imbalance issues were analyzed to improve the performance of the text classification algorithm. The overall accuracy of NB and KNN was 71.11% and 78.67%, respectively. The experimental results showed that KNN's performance was better than NB's [20]. Our study builds the news article classification model using LSTM+CNN, while the previous research builds the model using a machine learning algorithm.

- Wongso et al. [2] analyzed the suitable algorithm to classify news articles in Indonesian automatically. The dataset was retrieved using a web crawling method from www.cnnindonesia.com. The document first underwent several text preprocessing methods (i.e., lemmatization and stop word removal), followed by applying feature selection (i.e., term frequency-inverse document frequency (TF-IDF) and singular value decomposition (SVD) algorithms) and classification algorithms for Multinomial Nave Bayes, Multivariate Bernoulli Naïve Bayes, and Support Vector Machine. The test results suggested that the combination of TF-IDF and Multinomial Naïve Bayes Classifier gave the

highest results compared to the other algorithms, with a precision of 0.9841519 and a recall of 0.9840000. The results outperformed previous similar studies that classified Indonesian-language news articles with an accuracy of 85% [2]. The previous research builds the model using a machine learning algorithm, while the current study builds the news article classification model using LSTM+CNN.

- Fagbola et al. [4] evaluated the accuracy and efficiency of the computational time of Kolmogorov Complexity Distance Measure (KCDM) and Artificial Neural Network (ANN) for large-dimensional news article classification problems. Dataset used by British Broadcasting Corporation (BBC) News. The dataset consisted of 2225 news articles in five categories: politics (417), sports (511), entertainment (386), education and technology (401) and business (510). Porter's algorithm was used for words stemming after tokenization and deletion of stop words, and Normalized Term Frequency–Inverse Document Frequency (NTF-IDF) was adopted for feature extraction. Experimental results showed that ANN performed better in terms of accuracy while KCDM produced better results than ANN in terms of computational time efficiency [4]. The similarity of the previous study with the current study is that both utilized BBC dataset to evaluate the model. Nevertheless, the algorithm used in the feature extraction is different.
- Sunagar & Kanavalli [21] dealt with the complexities involved in the text classification process. The experiment was carried out with the implementation of the RNN+LSTM+Gated Recurrent Unit (GRU) model. This model was compared with RCNN+LSTM and RNN+GRU. The model was tested using the GloVe dataset. The accuracy and recall obtained from the models were assessed. The F1 score was used to compare the performance of the two models. The hybrid RNN model had three LSTM layers and two GRU layers, while the RCNN model contained four convolution layers and four LSTM levels, and the RNN model contained four GRU layers. The weighted average for the hybrid RNN model was found to be 0.74, RCNN+LSTM was 0.69, and RNN+GRU was 0.77. The RNN+LSTM+GRU model showed moderate accuracy in the initial epoch, but the accuracy slowly increased as the epoch increased. In contrast, the research in the article builds the news article classification model using LSTM+CNN.

Several related studies (review papers) on text classification suggest that machine learning algorithms are generally used for text classification combined with word embedding models. Text classification using a deep learning method that combines RNN/LSTM and CNN is still limited. In this study, we proposed a hybrid modeling (RNN/LSTM + CNN) with a word embedding feature (i.e., Word2vec + LSA/SVD) for news text classification and coupled with a model for outlier detection using machine learning algorithms, especially the decision tree algorithm.

III. Methodology

The hybrid model developed in this study consisted of two 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 was used to classify news articles based on Word2Vec-based context and LSA/SVD-based content. The outlier detection model was intended to predict which news was an outlier. Before the dataset was ready to be used for modeling, the dataset needed 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. After the data was ready to be used, the data was used for model training, which was then employed for data testing to validate the prediction results of the two models. In general, the hybrid model developed is presented in Fig. 1.

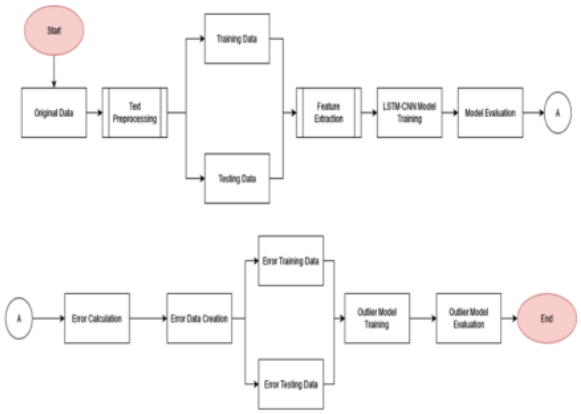

Fig. 1. Hybrid model framework

The classification results from the LSTM+CNN model was used as input in the outlier detection process. There would be errors (outliers) from the classification results generated by the model. For example, if a news item had topic A, the model might predict the probability that the news would be classified as topic B. Specifically, the use of the term "topic" in the research referred to content. Later, each news item would be labeled as an outlier or not an outlier. Furthermore, if topic A is at number 0.4, then there will be an error of 0.6. Once the error calculation result was obtained, the root mean square error (RMSE) for the entire dataset could be calculated. Sample/raw data that had an error of more than the RMSE would be categorized as an outlier. The outlier labeling enabled us to build an outlier classification model.

A. Text Preprocessing

The stages of text preprocessing news articles from raw data to ready-to-use data consisted of four stages (as presented in Fig. 2), namely deletion of symbols and numbers, tokenization, deletion of stop words, and lemmatization. The removal of symbols and numbers was done because symbols and numbers did not have a special meaning that was correlated with the topic of the news. Then, tokenization was done to break sentences or paragraphs in the news into chunks of words so that the model could read them. Furthermore, words with no meaning, such as subject (I, he, she, etc.), were

removed to reduce noise in the data. Finally, lemmatization was used to reduce the word form to its simplest form, such as "eating" to "eat".

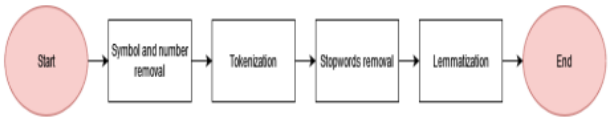

Fig. 2. Stages of text preprocessing

B. Feature Extraction

The news text that had been cleaned was not yet fully usable by the model. Feature extraction was used to extract information from the text. The feature extraction stages are presented in Fig. 3. This research used two extraction methods, namely context and content-based information/news extraction. The Word2Vec model was used to extract context-based information. Each word from the text processing above was converted into a 100-dimensional vector. A window of value 5 was used to obtain the vector, meaning that five surrounding words would be used to understand the context of a word in the paragraph. The results of the Word2Vec model were vectors with dimensions of 100, with the number of vectors as many as the number of words contained in the training data.

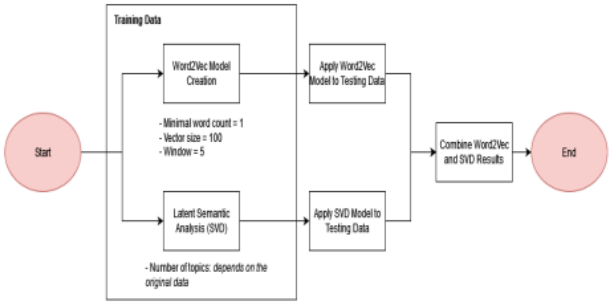

Fig. 3. Stages of feature extraction

The method used to extract content-based news was the SVD model. This model broke down the matrix of words obtained from the text processing results into three matrices, namely a matrix containing the relationship between news and words, a matrix containing the relationship between news and topics, and information containing the relationship between topics and words. The matrix used was the relationship matrix between topics and words so that each word had a vector containing information about its relationship to the topics. After obtaining two types of vectors from the results of Word2Vec and SVD, the two outputs were combined into one vector. This vector was called the embedding vector, which represented one word.

C. Hybrid Model of LSTM+CNN Architecture

Fig. 4 presents the architecture of the classification model of the Hybrid model using two deep learning algorithms (i.e., RNN/LSTM+CNN).

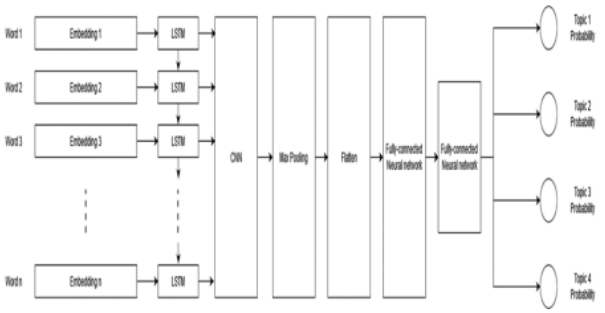

Fig. 4. Hybrid model architecture

Embedding vector from feature extraction was used in LSTM-CNN modeling as an embedding layer. When there was news used as model input, every word in the news would be transformed into the corresponding embedding vector. Then, the embedding vector became the LSTM input layer. The LSTM cells used were 64 cells. The output of each LSTM cell would be used as the CNN input layer, with a filter size of 100, a kernel size of 2, and a ReLu activation function. The CNN output matrix was inserted into the max pooling layer to reduce noise in the output. After the noise in the output was reduced, the output, which had been converted into a 1-dimensional vector in the flattened layer, can be input to a fully connected neural network (NN). Here, we use 2 NN layers, where the first layer had 16 nodes and the ReLu activation function, while the second layer had the number of nodes corresponding to the number of topics in the data. The softmax activation function was used in the second layer to obtain the output in the form of the probability of each topic (content).

D. Model Evaluation

The ROC curve provides a graphical representation of the performance of the classifier. The ROC curve was generated by calculating and plotting the TPR against the false positive rate (FPR) for a single classifier at various thresholds. The TPR and FPR equations are presented in equations 1 and 2.

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

$$FPR = 1 - Specificity = FP/(FP + TN) \quad (2)$$

Where TP is the number of true positives and FN is the number of false negatives. TPR is a measure of the probability that an actual positive event will be classified as positive. FP is the number of false positives and TN is the number of True Negatives. FPR is a measure of how often a "false alarm" will occur or how often an instance of a true negative will be classified as positive.

Visualization of ROC curve uses AUC. The higher the AUC score, the better the classifier performs for a given task. ROC curve with AUC score is presented in Fig. 5.

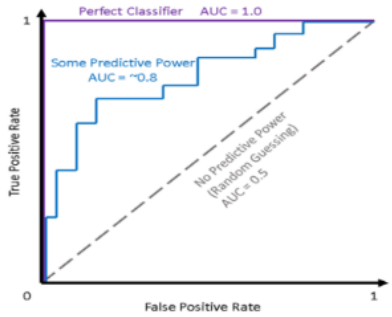

Fig. 5. ROC curve with AUC score

The classification category based on the ROC Curve value is shown in Table I [22]. In general, an AUC score of 0.6 - 0.7 indicates poor classification performance (failure), 0.7 - 0.8 is considered acceptable, 0.8 - 0.9 is considered very good and more than 0.9 is considered extraordinary. Model category based on ROC curve value [23] is presented in Table I.

TABLE I. MODEL CATEGORY BASED ON ROC CURVE

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

IV. EXPERIMENTAL RESULT

A. News Classification

The architecture used to build the model with the AGNews dataset consisted of: (1) Embedding matrix (the result of extracting content and context information that was carried out in the previous stage) with input size = 100; (2) LSTM with 64 nodes = 64; (3) CNN with parameters: filter = 100, kernel = 2, number of strides = 1, padding = valid, activation function = Relu; (4) Pooling; (5) Flatten; (6) Dense hidden layer with number of nodes = 16 and activation function = Relu; (7) Dropout rate = 0.5; (8) Dense hidden layer with the number of nodes = 4 (according to the number of targets) and the activation function = softmax. Meanwhile, the hybrid architecture model for the BBC News dataset, which were points 1 to 7, was the same as the AGNews dataset. The difference between the two dataset was in point 8. The dense hidden layer was used in BBC News with five nodes (according to the number of targets). This step was done to achieve the best set of hyperparameters by performing hyperparameter tuning.

The next step was the model fitting. This stage conducted model training, which would be stopped if there was no increase in the accuracy value in data testing for 10 iterations (epochs). The modeling results were then stored (Table II).

TABLE II. HYPER PARAMETERS

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

After the hybrid model was built using training data with hyperparameter tuning, the model was evaluated using the confusion matrix (CM) and ROC curve. The test results after 235 iterations showed a loss value of 0.2818, accuracy of 0.9065, test loss of 0.28183448, test accuracy of 0.90653336. The CM on the model with the AGNews dataset suggested that the data had been predicted correctly for each actual label, as shown in Table III.

TABLE III. 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 CM on BBC News is presented in Table IV.

TABLE IV. 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 was 91%, the ROC curve value was 0.9832 and the training validation had a loss value of 0.1146 and an accuracy of 0.9611 (96.11%). The results of accuracy in training, testing and validation as well as the ROC curve value suggested that the model had good performance. The ROC curve for AGNews is presented in Fig. 6.

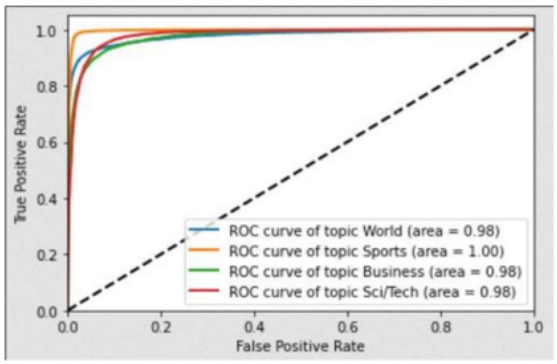

Fig. 6. ROC curve of the AGNews dataset

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

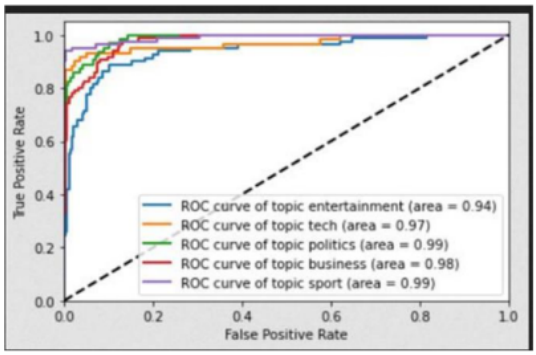

Fig. 7. ROC curve of the BBC News dataset

The ROC curve showed the visualization between TPR and FPR. The classifier that provides the curve closer to the top left corner (perfect classifier) exhibits better performance. The closer the curve is to the 45-degree diagonal of the ROC space, the less accurate the classifier is.

TABLE V. 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 was trained and tested against two datasets, namely the AGNews and BBC News datasets. The results of the model evaluation showed that the LSTM+CNN hybrid model had excellent accuracy (as presented in Table V) and ROC AUC scores.

B. Outlier Detection

The first step in outlier detection is processing error data. At this stage, the prediction results of the error model will be collected into 1 (one) separate data frame consisting of test topics and train topics. The magnitude of the error is calculated through the equation: error = 1 – max(xi), i∈{1,2,3,4}error = 1 – max(xi), i∈{1,2,3,4}, where x_i denotes the highest probability score that the model assigns. In this formulation, each prediction error can be calculated by the magnitude of the error, so that the RMSE error data can be calculated. The size of the RMSE dataset error was 0.93. Fig. 8 shows error data train, while Fig. 9 shows error data test on the AGNews dataset.

	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
76249	[caetano, defender, sergirnho, 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 with an error greater than the RMSE was around 65% and this data would be considered as an outlier. Fig. 10 shows 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
135	[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

To predict whether a news item is an outlier, two types of systems were designed. In the first system (Fig. 11), the outlier classifier model was installed separately from the LSTM+CNN hybrid model. As for the second system (Fig. 12), the meta-modeling principle was used, namely the outlier

classifier model using the prediction results of another model (the LSTM+CNN hybrid model for this case) to determine whether a news item is an outlier news or not. Decision tree models were used in both systems to determine which system had the best performance in determining outliers.

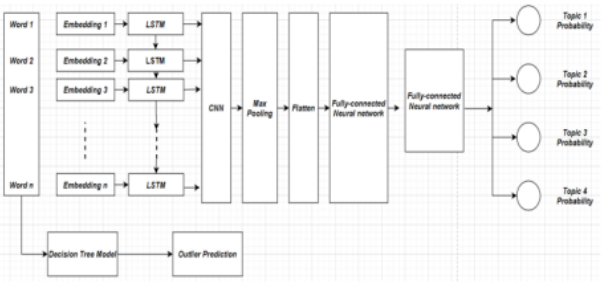

Fig. 11. The First Model of the classification of news outliers

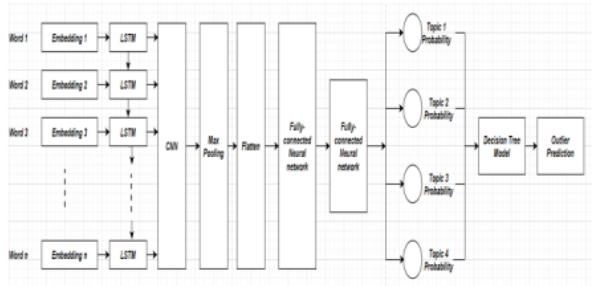

Fig. 12. The second model of the classification of news outliers

C. Validation Model

Validation models of the two classification models are presented in Fig. 13. Fig. 13 shows that system 2 is far superior to system 1. In system 1, the ROC AUC score was only around 0.5-0.6, meaning that the model predictions were not much different from the random prediction results. However, in system 2, the ROC AUC score could reach 0.8 - 0.9, indicating that by obtaining input from the probability prediction results of the LSTM-CNN hybrid model, the outlier classifier model can properly determine whether a news item is an outlier news or not.

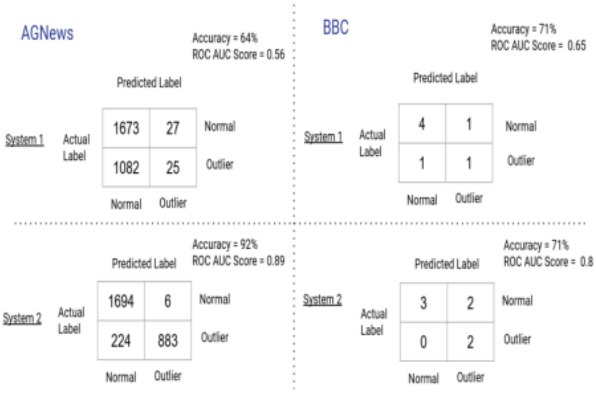

Fig. 13. Comparison of outlier news classifications

D. Comparative Study

Table VI is a comparison of previous research related to news article classification. The results of the performance evaluation of training and testing of the developed hybrid model showed better results. Based on the ROC curve, the hybrid model developed is in the excellent classification category. This can prove that the hybrid model with RNN/LSTM+CNN architecture and feature extraction Wor2Vec+LSA/SVD can be utilized as a good method for classifying sequential text. Furthermore, the hybrid model developed can perform news outlier detection well. The last two approaches in Table VI are the current research results (presented in bold).

TABLE VI. COMPARISON WITH PREVIOUS RESEARCH

Word Embedding	Text Classification Strategies	Dataset	Result
GloVe [21]	• RNN + LSTM + GRU • RCNN+ LSTM • RNN+GRU	GloVe	Accuracy = 0.74 Accuracy = 0.69 Accuracy = 0.77
GloVe, word2vec, and fastText [12]	CNN, SVM, XGBoost	RCV1	Precision value = 0.92, Recall = 0.92 and F1 measure = 0.920
Word2Vec, doc2vec, BoW [13]	SVM	VINCI	AUC value between 0.80 -0.93 and accuracy value between 0.82-0.86
BoW [14]	LSTM2: repLSTM and rankLSTM	Bio (10C), email and News	F1 Measure less than 75%
Word2Vec +LDA [9]	CNN	Sogou Corpus text classification Lab	Accuracy = 0.84, recall = 0.89 and f1-score = 0.86
Word2Vec + TF-IDF [15]	Att-LSTM	IMDB film review and hotel review corpora	Precision value = 0.87, recall = 0.87 and F1 measure = 0.87
[24]	Label-Embedding Attentive Model (LEAM)	AGNews, Yelp Review Full, Yelp Review Polarity, DBPedia and Yahoo! Answers Topic	F1 micro average = 0.91 and macro average = 0.88 under the ROC (AUC) curve, precision = 0.61
LDA + Word2Vec [17]	SVM	Corpus News	Precision value = 83.6, F1-score = 84.4 and recall = 85.4.
LDA [18]	Skip-Gram and CNN	Biomedical literature,	F1 score = 82.7%.

		clinical record annotation	
TF-IDF [19]	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 [20]	Naïve Bayes (NB) and K-Nearest Neighbor (K-NN)	Corpus Dataset	Naïve Bayes accuracy is 71.11%, and KNN is 78.67%.
TF-IDF dan SVD [2]	Multinomial Nave Bayes, Multivariate Bernoulli Naïve Bayes, and Support Vector Machine	CNN indonesia	TF-IDF and Multinomial Naïve Bayes Classifier provide the highest results compared to other algorithms, with a precision of 0.9841519 and a recall of 0.984
Normalized TF-IDF (NTF-IDF) [4]	Kolmogorov Complexity Distance Measure (KCDM) and ANN	British Broadcasting Corporation (BBC) News	ANN performs better in terms of accuracy while KCDM produces better results than 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: Accuracy = 0.71 and ROC curve = 0.8**

V. CONCLUSION

In this study, we developed a hybrid model consisting of two models: the news classification model (news categories) and the outlier classification model. The news classification model employed a deep learning algorithm (i.e., LSTM+CNN) based on the context and content of a news story. Meanwhile, the outlier classifier model was intended to predict which news is an outlier. The datasets used for modeling were AGNews and BBC News. The Word2Vec model was used to extract context- and content-based news using the SVD model. Our results on AGNews showed an accuracy of 0.91 with a ROC curve score of 0.97, while BBC News had an accuracy value of 0.86 with a ROC curve score of 0.96. These results suggested that the hybrid model had excellent accuracy and ROC AUC scores.

The process of labeling news included the detection of outliers. The predicted data from the LSTM-CNN model was used and tested with two types of models. First, the outlier classifier model was installed separately from the LSTM-CNN hybrid model. The second model used the meta-modeling principle, namely the outlier classifier model using the prediction results of another model (LSTM-CNN hybrid model). The algorithm used in the outlier classification model was a decision tree. Of the two models tested, the second model was far superior to the first model. In the first model, the ROC AUC score was only around 0.5-0.6, which indicated that the model predictions were not much different from the random prediction results. However, in the second model, the ROC AUC score could reach 0.8 - 0.9, meaning that by obtaining input from the probability prediction results of the LSTM-CNN hybrid model, the outlier classifier model could properly determine which news item was an outlier, with accuracy values of 0.71 and 0.92 for BBC News AGNews, respectively.

The drawback of the results of this study is that the proposed hybrid model did not reach maximum accuracy. The sample used in the dataset to evaluate the performance of the model is still limited, causing the model to have difficulties learning the vocabulary used in classifying news articles based on their categories. This situation may lead to overfitting. Furthermore, the model only tested news articles in English. Further research is needed to develop the proposed hybrid model, combined with other deep learning algorithms and more dataset samples, to obtain optimum model performance. The performance of the developed model also needs to be tested using news articles other than in English, such as Indonesian, Mandarin, Arabic and other languages.

REFERENCES

[1] D. Liparas, Y. HaCohen-Kerner, A. Moumtzidou, S. Vrochidis, and 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, pp. 63–75, October 2014, doi: 10.1007/978-3-319-12979-2_6.

[2] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, and Rudy, "News article text classification in Indonesian language," Procedia Comput. Sci., vol. 116, pp. 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, pp. 2555–2562, 2020.

[4] T. M. Fagbola, C. S. Thakur, and O. Olugbara, "News article classification using Kolmogorov Complexity Distance Measure and Artificial Neural Network," Int. J. Technol., vol. 10, no. 4, pp. 710–720, 2019.

[5] S. Biradar and M. M. Raikar, "Performance analysis of text classifiers based on news articles-a Survey," Indian J. Sci. Res., vol. 15, no. 2, pp. 156–161, 2017.

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

[7] W. K. Sari, D. P. Rini, and R. F. Malik, "Text classification using Long Short-Term Memory with GloVe features," J. Ilm. Tek. Elektro Komput. dan Inform., vol. 5, no. 2, pp. 85, 2020, doi: 10.26555/jiteki.v5i2.15021.

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

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

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

[11] P. Krammer, O. Habala, J. Mojžiš, L. Hluchý, and M. Jurkovič, "Anomaly detection method for online discussion," Procedia Comput. Sci., vol. 155, pp. 311–318, 2019, doi: https://doi.org/10.1016/j.procs.2019.08.045.

[12] A. J. Stein, J. Weerasinghe, S. Mancoridis, and R. Greenstadt, "News article text classification and summary for authors and topics," pp. 1–12, 2020, doi: 10.5121/csit.2020.101401.

[13] Y. Shao, S. Taylor, N. Marshall, C. Morioka, and 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, pp. 2874–2878, 2019, doi: 10.1109/BigData.2018.8622345.

[14] Y. Yan, Y. Wang, W. C. Gao, B. W. Zhang, C. Yang, and X. C. Yin, "LSTM 2: multi-label ranking for document classification," Neural Process. Lett., vol. 47, no. 1, pp. 117–138, 2018, doi: 10.1007/s11063-017-9636-0.

[15] H. Yuan, Y. Wang, X. Feng, and S. Sun, "Sentiment analysis based on weighted Word2vec and ATT-LSTM," ACM Int. Conf. Proceeding Ser., pp. 420–424, 2018, doi: 10.1145/3297156.3297228.

[16] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and 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, pp. 1101–1111, 2018, doi: 10.18653/v1/n18-1100.

[17] F. Sun and 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.

[18] H. Xu, A. Kotov, M. Dong, A. I. Carcone, D. Zhu, and 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, pp. 88–97, 2016, doi: 10.1145/2975167.2975176.

[19] M. Goudjil, M. Koudil, M. Bedda, and N. Ghoggali, "A novel active learning method using SVM for text classification," Int. J. Autom. Comput., vol. 15, no. 3, pp. 290–298, 2018, doi: 10.1007/s11633-015-0912-z.

[20] M. Azam, T. Ahmed, F. Sabah, and M. I. Hussain, "Feature extraction based text classification using K-Nearest Neighbor Algorithm," IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 18, no. 12, pp. 95–101, 2018, [Daring]. Available on: http://paper.ijcsns.org/07_book/201812/20181213.pdf.

[21] P. Sunagar and A. Kanavalli, "A hybrid RNN based deep learning approach for text classification," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 6, pp. 289–295, 2022.

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

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

[24] G. Wang et al., "Joint embedding of words and labels for text classification," ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 1, pp. 2321–2331, 2018, doi: 10.18653/v1/p18-1216.

Hybrid Modeling to Classify and Detect Outliers on Multilabel Dataset based on Content and Context

ORIGINALITY REPORT

0%
SIMILARITY INDEX

PRIMARY SOURCES

1	journal.yrpipku.com Internet	9 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF