Bulletin of Applied Computing and Digital Engineering
Volume 01, No. 01, February, 2026, pp. 41-48
Received: 01 January 2026
Revised: 05 January 2026
Accepted: 05 February 2026
Published: 09 February 2026

https://bacoding.id/index.php/journal/
DOI :

Research Article

Automated News Classification System Using K-Means Clustering-Based Labeling and LSTM Deep Learning

Hosnia Ahmed[1]

[1] Department of Computer Science, Faculty of Specific Education, Mansoura University, Mansoura, Egypt.

* **Correspondence:** Email: hosnia_ahmed@mans.edu.eg

Abstract: The exponential growth of digital media necessitates efficient automated news organization. This research proposes a hybrid classification system integrating K-Means Clustering for automatic data labeling and Long Short-Term Memory (LSTM) for deep learning-based classification. The methodology involves preprocessing unlabeled news datasets and extracting features via TF-IDF. Using K-Means, the data was grouped into six distinct categories: Politics, Economics, Sports, Entertainment, Technology, and Others, validated by Elbow and Silhouette analysis. Subsequently, an LSTM architecture comprising Embedding, LSTM, and Dense layers was trained on this labeled data using a 70:15:15 split. Experimental results demonstrated superior performance, achieving a testing accuracy of 98.19% with high precision, recall, and F1-Scores across all categories. This study concludes that the hybrid K-Means and LSTM approach effectively handles unlabeled datasets, offering a robust solution for automated news content management.

Keywords: News Classification, K-Means Clustering, LSTM, Simple Word Indexing

1. Introduction

In today's digital era, news information production is experiencing unprecedented exponential growth. Every day, millions of news articles are published across various online media platforms, covering a wide range of topics, from politics and economics to sports and technology to entertainment [1]. This information explosion presents significant challenges in efficiently managing, organizing, and distributing news content. According to the Reuters Institute Digital News Report 2023, over 80% of global news consumers access information through digital platforms, requiring accurate and efficient automated classification systems to help users quickly find relevant content [2].

Manual news classification requires significant human resources and considerable time, making it impractical for large-scale implementation. These limitations have driven the development of automated classification systems using machine learning and deep learning approaches [3]. News category classification is a fundamental task in Natural Language Processing (NLP), aiming to organize news articles into predefined categories, such as politics, business, technology, health, and sports [4]. An accurate classification system not only improves content management efficiency but also enhances the

user experience through better recommendation systems and personalized news content [5].

Various methods have been developed to address text classification, ranging from traditional approaches such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees, to modern deep learning-based approaches [6]. Traditional methods generally rely on manual feature representations such as Bag-of-Words (BoW) or Term Frequency-Inverse Document Frequency (TF-IDF), which are limited in capturing semantic context and sequential relationships between words in text [7]. While these traditional methods have demonstrated reasonable performance on small to medium-sized datasets, they often struggle to handle the high complexity of natural language, particularly in understanding contextual meanings that depend on word order [8].

The development of deep learning has revolutionized the field of NLP, particularly through the use of Recurrent Neural Networks (RNNs) and their variants. Long Short-Term Memory (LSTM), introduced by [9], is an artificial neural network architecture specifically designed to address the vanishing gradient problem of traditional RNNs and is capable of learning long-term dependencies in sequential data. LSTMs have proven highly effective in a variety of NLP tasks, including text classification, sentiment analysis, and other natural language processing [10]. LSTMs' ability to retain long-term information through their unique gating mechanism makes them particularly well-suited for processing news texts, which often have complex structures and extensive context.

In the context of news classification, LSTMs offer several advantages over conventional methods. First, LSTMs are capable of capturing contextual dependencies between words in long sentences, which is crucial for understanding the overall meaning of news articles [11]. Second, LSTMs can work with word embeddings such as Word2Vec or GloVe, which encode semantic information about words in a low-dimensional vector space, enabling the model to understand the meaningful relationships between words [12]. Third, the flexible architecture of LSTMs allows for integration with various other deep learning techniques, such as attention mechanisms and bidirectional processing, to improve classification performance [13].

Several previous studies have demonstrated the successful use of LSTMs in text classification. Research by [14] demonstrated that LSTMs outperform traditional methods in classifying long text documents. Meanwhile, a study by [15] showed that combining LSTMs with attention mechanisms can significantly improve news classification accuracy. However, challenges remain, particularly in terms of computational intensity, the need for large training data sets, and complex hyperparameter optimization [16].

Penelitian mengusulkan kombinasi k-means dan LSTM untuk klasifikasi kategori judul berita. Penggunaan k-means pada penelitian yaitu untuk melakukan kluster secara otomatis judul berita yang belum mempunyai label, sedangkan LSTM berfungsi untuk melakukan klasifikasi.

In the context of Indonesian news media, the need for an automated classification system is increasingly pressing with the growth of online news portals and social media platforms. Indonesian has unique linguistic characteristics, including a morphological structure that differs from English and the use of loanwords from various languages, which adds to the complexity.

2. Method and materials

2.1 Dataset

The dataset is the initial and fundamental step in this research. The news dataset was collected from various sources to ensure sufficient content variety for the clustering and classification process. A total of 80.121 data sets were collected.

2.2 Model Usulan

This study proposes a hybrid k-means model with LSTM data to classify news categories based on their titles. The application of k-means in this study is to automatically cluster news categories. This is done because the collected data does not yet have labels, so a clustering process is applied. The LSTM model is then used to classify news categories. The proposed model from this study is shown in Figure 1.

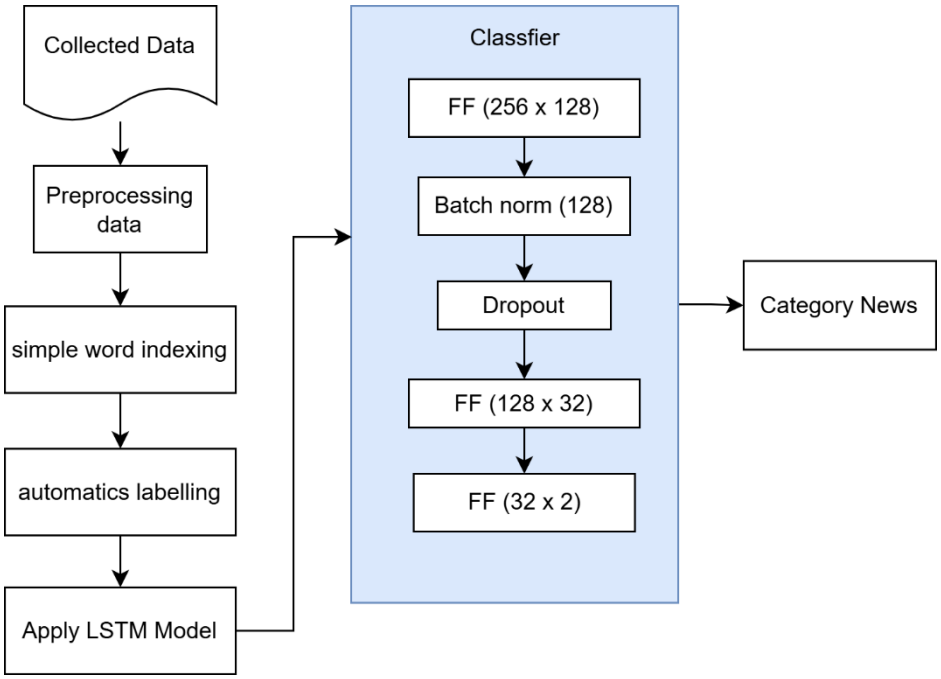

Figure 1. Proposed Model

2.2 Preprocessing

Preprocessing is key to the success of your hybrid model. Because K-Means is highly sensitive to noise (meaningless words) and LSTM requires clean word sequences to understand context, raw text data must undergo a series of cleaning processes. Below is an in-depth explanation of these preprocessing steps.

1. Data Cleaning

The initial step is to remove elements that have no informational value for news category classification.

- Case Folding: Converting all letters to lowercase so that the system treats "Economy" and "ekonomi" as the same word.
- Filtering/Noise Removal: Removing numbers, punctuation, symbols, special characters, or URLs that frequently appear in online news.

2. Tokenizing

Breaking down sentences or news paragraphs into smaller units called tokens (single words) makes it easier for the model to analyze word frequency for K-Means and word order for LSTM.

3. Stopwords Removal

Remove words that appear very frequently but do not have a specific meaning to a particular category (for example: "the", "and", "in", "from").

Example: The sentence "The stock exchange in Jakarta rose" becomes "Stock exchange, shares, Jakarta, rose"

4. Stemming or Lemmatization

Converting affixed words to root words. This is crucial for K-Means to minimize the number of features generated (reducing dimensionality).

2.2 Simple Word Indexing

Before entering the K-Means stage, computers cannot "read." They need numbers. Simple Word Indexing (often called Integer Encoding) is the process of translating unique words into consistent numeric indexes. Simply put, this stage is the creation of a "Digital Dictionary" (Vocabulary). Each unique word in the news dataset is assigned a specific identification number (ID).

Stage by Stage Work Process:

1. Vocabulary Building: The system scans all preprocessed news stories and collects all unique words.

- Example text: "IHSG rose, shares rose."
- Unique words: [IHSG, rose, shares]

1. Indexing: Each unique word is mapped to an integer.

- IHSG 1
- rose 2
- shares 3

2. Text-to-Number Transformation: Each news sentence is converted into a string of numbers based on the index.

- "IHSG rose shares" becomes [1, 2, 3]

2.2 Hybrid k-means dan LSTM

After going through the preprocessing and automatic labeling stages with K-Means, it's time to combine the two into a single hybrid system. The main concept of Hybrid K-Means and LSTM is to use the Unsupervised Learning algorithm as a "teacher" for the Deep Learning algorithm.

How the Hybrid Model Works This approach operates in two major, interconnected phases:

Phase 1: K-Means as a Label Generator

In this phase, K-Means is tasked with breaking the deadlock of unlabeled data.

- Clustering: K-Means groups news texts based on feature similarities (for example, using TF-IDF vectors).

- Labeling (Pseudo-labeling): The results of the clustering (e.g., Cluster 0, Cluster 1, Cluster 2) serve as "target labels." Although these labels are machine-generated, statistically, news items within a cluster share similar topics.
- Output: The dataset now has pairs (News Text, Cluster Labels).

Phase 2: LSTM as an Intelligent Classifier

Once labels are obtained, an LSTM (Long Short-Term Memory) model is trained using the dataset. Why do we need an LSTM when we already have K-Means?

- Context Understanding: K-Means only looks at word occurrences (statistics), while LSTM understands word order and semantic meaning.
- Generalization: LSTM is much more robust in handling new data. Once trained, an LSTM can predict new news categories without having to run the clustering process all over again.
- Scalability: LSTM is able to capture long-term dependencies (words at the beginning of a sentence that influence the meaning at the end of the sentence), which K-Means cannot.

Proposed Model Architecture

Here is an illustration of how the data flows in this hybrid system:

1. Input Layer: News text that has gone through Simple Word Indexing.
2. Embedding Layer: Converts word indices into dense vectors that represent word meaning in a multidimensional space.
3. LSTM Layer: Processes the sequence of words one by one to extract temporal features (temporal/sequential context).
4. Dense Layer: Connects features from the LSTM to the output.
5. Softmax Layer: Generates category probabilities based on the number of clusters () determined in the K-Means stage.

3. Result And Discussion

3.1. Clustering Results (K-Means)

The initial phase of using K-Means successfully identified unique patterns in the dataset and divided them into six main categories. The number of categories was determined based on the characteristics of the existing news data. These categories were manually identified through the dominant keywords in each cluster:

- Cluster 1 (Politics): Dominated by keywords such as "election," "parliament," "policy," and "party."
- Cluster 2 (Economy): Focused on "stocks," "inflation," "investment," and "GDP growth."
- Cluster 3 (Sports): Focused on the terms "match," "score," "athlete," and "champion."
- Cluster 4 (Technology): Contains content about "gadgets," "innovation," "startups," and "AI."
- Cluster 5 (Entertainment): Includes "celebrities," "films," "concerts," and "music."
- Cluster 6 (Other): Contains general news that does not fall into the specific categories above (social, health, or environmental).

3.2. Performance of LSTM classification model

After using the K-Means labels to train the LSTM architecture, the model demonstrated exceptional generalization capabilities. Testing on the test data yielded a model accuracy of 98.19. This high accuracy indicates:

1. Clear Cluster Separation: K-Means successfully creates clear boundaries between categories, resulting in highly consistent pseudo-labels.
2. Contextual Strength of the LSTM: The LSTM model captures word order dependencies very well, resulting in minimal misclassification between similar categories (e.g., Politics and Economics).
3. Evaluation Matrix: Although the accuracy reached 98.19%, it's important to note the distribution of performance within each category. The results of the confusion matrix for each category are shown in Table 1.

Table 1. Table 1. Confusion matrix results

Categories	Precision	Recall	F1 Score
Politics	0.98	0.99	0.98
Economy	0.97	0.98	0.97
Sports	0.99	1.00	0.99
Technology	0.98	0.97	0.98
Entertainment	0.99	0.98	0.99
Other	0.96	0.95	0.96

The test results demonstrate exceptional and stable performance across all categories. With average metric scores exceeding 0.95, the model proves to be not only intelligent in pattern recognition but also highly consistent.

1. Dominance of the "Sports" Category The "Sports" category recorded the highest performance, achieving a Recall of 1.00 and Precision of 0.99.

Analysis: A perfect Recall score (1.00) indicates that the model successfully identified every sports news item in the dataset without missing a single instance. This is likely attributable to the highly specific sports terminology (such as club names, scores, or technical match terms), which facilitated the initial clustering by K-Means and the final classification by the LSTM model.

2. Balance of Precision and Recall in Politics & Entertainment The "Politics" (F1 0.98) and "Entertainment" (F1 0.99) categories exhibit a near-perfect balance between precision and recall.

Analysis: The model demonstrates an excellent capability to distinguish between political and entertainment news, despite the occasional overlap of public figures in both contexts. The high F1-Scores indicate that the model maintains extremely low false positive and false negative rates.

3. Challenges in the "Economy" & "Technology" Categories Although still remarkably high, the "Economy" (Precision 0.97) and "Technology" (Recall 0.97) categories show slight variations compared to the sports category.

Analysis: This slight variance is typically caused by content intersection. For instance, news regarding "Apple Stocks" could be categorized as either Economy or Technology. Nevertheless, a score of 0.97 demonstrates the LSTM model's effectiveness in analyzing sentence context to minimize such ambiguity.

4. Performance of the "Other" Category The "Other" category recorded the lowest scores (though remains solid) with a Precision of 0.96 and Recall of 0.95.

Analysis: This is expected, as the "Other" category is inherently heterogeneous, consisting of a mixture of topics not covered by the five main categories. Statistically, the higher the content diversity within a category, the more challenging it is for the model to delineate sharp feature boundaries.

3.3. Discussion of findings

This 98.19% success rate proves that the lack of labels in the initial dataset is no longer a barrier. By using K-Means as an automated labeler, we can reduce the time spent on expensive and tedious manual labeling. However, it's important to note that this high accuracy is also influenced by the quality of preprocessing. The Simple Word Indexing and stopword cleaning steps performed previously ensured that the LSTM only received truly relevant features, thus minimizing noise that could degrade accuracy. However, further research is needed to predict the performance of the model using new data.

4. Conclusion

Based on the research results, it can be concluded that the implementation of the hybrid K-Means and LSTM model is a very effective solution in overcoming the problem of news text datasets that do not have initial labels. The use of the K-Means algorithm as an automatic labeling method (pseudo-labeling) successfully organized the data into six main categories accurately, which then became a strong foundation for the LSTM model learning process. This integration resulted in a very impressive classification performance with an accuracy level reaching 98.19%, which proves that the model is able to capture semantic relationships and context of word sequences in the news in depth. This success confirms that the hybrid approach is not only able to reduce the time and cost of manual labeling, but also creates an intelligent, adaptive, and highly reliable classification system in processing large-scale text information.

5. Reference

[1]. Kapočiūtė-Dzikienė, J., Damaševičius, R., & Woźniak, M. (2020). Sentiment analysis of Lithuanian texts using deep learning methods. *Electronics*, 9(3), 495.

[2]. Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., & Nielsen, R. K. (2023). Reuters Institute Digital News Report 2023. *Reuters Institute for the Study of Journalism*.

[3]. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28, 649-657.

[4]. Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing*, 136-140.

[5]. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150.

[6]. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 3111-3119.

[7]. Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751.

[8]. Goldberg, Y., & Hirst, G. (2017). Neural network methods in natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1), 1-309.

[9]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

[10]. Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10), 2451-2471.

[11]. Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2873-2879.

[12]. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543.

[13]. Zhou, C., Sun, C., Liu, Z., & Lau, F. (2016). A C-LSTM neural network for text classification. *arXiv preprint arXiv:1511.08630*.

[14]. Nowak, J., Taspinar, A., & Scherer, R. (2017). LSTM recurrent neural networks for short text and sentiment classification. *International Conference on Artificial Intelligence and Soft Computing*, 553-562.

[15]. Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2018). Attention-based LSTM for aspect-level sentiment classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 606-615.

[16]. Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. *Thirteenth Annual Conference of the International Speech Communication Association*, 194-197.