IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2024

DOI: 10.17148/IJARCCE.2024.131260

Implementation of Data Retrieval Model Based on Semantic Similarity Analysis using Deep Learning Application

Ankush R. Deshmukh [1]* Dr.P.B.Ambhore [2]*

Research Scholar, Government College of Engineering,Amravati[1]

ankudesh@gmail.com

Assistant Professor, Government College of Engineering,Amravati[2]

pbambhore@gmail.com

Abstract: This project aims to develop a deep learning-based text classification system that predicts the domain of a given article using the 20 Newsgroups dataset, which consists of news articles categorized into various topics. The goal is to classify articles into broader domains such as 'Technology,' 'Sports,' 'Politics,' and 'Religion,' based on their content. The model employs an LSTM network, a form of RNN, because it is well-suited to handle sequential data like text and capture long-term dependencies in the content. The project first preprocesses the data by tokenizing the text, padding sequences to have uniform input size, and one-hot encoding the target labels. Next, the LSTM network is trained so that it may recognize the text's patterns and features and be able to map it into a predefined category. The model was evaluated in terms of accuracy, precision, recall, and F1-score. Also, the batch size and number of epochs were readjusted according to hyperparameter tuning for increased accuracy. Through training, the model can predict the category of any unseen article. The result is mapped to its corresponding domain using a predefined dictionary. The system also maintains the functionality of saving the trained model, tokenizer, and label encoder so that the same model can easily be loaded for further predictions. This text classification system can be applied in areas such as news aggregation, content categorization, and information retrieval where automatic sorting of articles into relevant domains is required. In addition, the project explores the possibility of improving text classification using LSTM networks on domains with large unstructured text data, thereby contributing to the advancements in NLP and deep learning applications in real-world scenarios.

Keywords: Text Classification, Deep Learning, LSTM , 20 Newsgroups Dataset, Recurrent Neural Networks (RNN), Content Categorization, Tokenization Sequence Padding.

I.INTRODUCTION

With the growing volume of textual data across domains, it is becoming increasingly difficult to manually classify and categorize large amounts of content. For instance, news articles cover an enormous range of topics, from technology to politics, sports, and religion. Such content needs efficient methods for automatic categorization. Traditional keyword-based classification methods often fail in complex and diverse text. Machine learning, particularly deep learning techniques, plays a crucial role in this regard. Deep learning, in recent times, has rapidly advanced the frontiers of NLP, empowering machines to decode and analyze text with remarkable precision. Among all deep learning models, LSTM-based networks, belonging to the Recurrent Neural Network (RNN) family, have been spectacularly promising in text classification because they can recognize long-term dependencies and contextual relations within sequential data. This project is intended to design an LSTM-based model for text classification using the 20 Newsgroups dataset. The dataset comprises about 20,000 documents categorized into 20 different newsgroups, making it perfect for training a model that can predict the domain or category of new, unseen articles. The main aim of this project is to build an effective model that can automatically categorize text into broad domains such as 'Technology,' 'Sports,' 'Politics,' and 'Religion.'

The process starts with data preprocessing, which includes tokenization, padding sequences to ensure uniform input sizes, and encoding labels for the different categories. The LSTM network is trained to recognize the patterns and relationships within the text and predict its corresponding category. After training, the model's performance will be evaluated using standard metrics like accuracy, precision, recall, and F1-score. The system will also feature the ability to save and load the trained model for future predictions. In several areas, ranging from automated categorization of news to content recommendations and information retrieval, the application of the outputs of this project could be extensive.

DOI: 10.17148/IJARCCE.2024.131260

This capacity of the classification of articles-accurate as well as timely-can prove really useful for organizational management and huge volumes of text data. Along with this, the project explores the applicative realization of deep learning techniques in the NLP that shows the efficiency of LSTM for real-world tasks of text-classification.

II. LITERATURE REVIEW

Semantic similarity analysis has emerged as a crucial technique in improving the effectiveness of data retrieval systems. Traditional information retrieval methods, such as keyword-based searches (e.g., TF-IDF), often fall short when the focus shifts to understanding the meaning behind a query or document. These models rely heavily on exact keyword matches and fail to capture synonyms, context, and deeper semantic relationships between words. This limitation has led to the development of more advanced methods that aim to improve the semantic understanding of the text.

Author	Methodology	Identified Gap
[1] Agerri, R., & Garcia-Serrano, A. (2018).	Surveys rule-based, machine learning, and deep learning approaches for STS, including models like LSTM.	Need for models that generalize across domains without domain-specific fine-tuning.
[2] Wang, X., & Jiang, L. (2020).	Surveys deep learning techniques like Siamese networks, LSTMs, and BERT for text similarity tasks.	Difficulty in handling long-term dependencies and context effectively. Scaling models for large datasets.
[3] Chen, X., et al. (2019	Investigates the use of BiLSTM for semantic similarity prediction tasks, processing text in both directions.	Gap in integrating knowledge from various domains for improved prediction accuracy and reduced bias.
[4] Mikolov, T., et al. (2013)	Focuses on deep learning techniques (Word2Vec, GloVe) for learning word and sentence embeddings.	Need for models incorporating domain-specific knowledge while maintaining contextual understanding.
[5] Li, Z., & Yang, L. (2018)	Combines word embeddings (Word2Vec) with LSTM networks for semantic similarity prediction.	Need for models that address similarity for paraphrased or contextually altered text. Fine-tuning embeddings.
[6] Vaswani, A., et al. (2015)	Explores deep RNNs, including LSTM and GRU, for modeling text similarity by capturing temporal dependencies.	Shortcomings in handling varied sentence structures across different languages and domains.
[7] Jadhav, P., & Mishra, B. (2020).	Investigates various deep learning architectures (CNN, LSTM, BiLSTM) for sentence-level semantic similarity tasks.	Lack of universal architecture that works equally well across different domains and types of sentence pairs.

III. DATASET

The 20 Newsgroups dataset is one of the most popular resources in Natural Language Processing (NLP) and machine learning for text classification and topic modeling. It contains about 20,000 documents, which are posts from different newsgroups, categorized into 20 distinct topics. These topics include diverse areas such as technology, politics, sports, religion, and more. Some of the categories in the dataset include Comp.sys.ibm.pc.hardware (technology), Talk.politics.misc (politics), and Sci.space (science), among others.The dataset is divided into two sets: a training set that contains 11,314 documents, and a test set that contains 7,532 documents. The document for every post is the raw text coming from a newsgroup. Some basic headers, like the subject lines, are available, but nothing more in the way of timestamps or author names.

To prepare the dataset for machine learning, preprocessing steps are crucial. These typically include tokenization (splitting the text into words or subwords), stopword removal (eliminating common words like "the" and "is" that do not add meaning), lowercasing (standardizing case to treat words like "Computer" and "computer" as the same), and removing special characters (such as punctuation, URLs, or numbers).The 20 Newsgroups dataset is used in tasks such as text classification, where the model learns to predict the category of a document based on its content, topic modeling, where dominant themes in the text are identified, and sentiment analysis. Despite its widespread adoption, it is noisy text - for example, email headers and signatures - it has imbalanced class distributions in some categories that have more documents than others - and variability in document quality such as a low-quality post that needs to be addressed in the development and evaluation of the model. Nonetheless, it remains a good benchmark for testing text classification algorithms and machine learning models.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2024

DOI: 10.17148/IJARCCE.2024.131260

IV. OBJECTIVES

Objectives:

The objectives of work is as follows

Text Classification: This project aims at creating an efficient deep learning-based model that could classify text articles into predefined categories using the 20 Newsgroups dataset. The model predicts the category of the article to be 'Technology,' 'Sports,' 'Politics,' or 'Religion,' based on the semantic content of the text.

Model Development: To develop and train a Long Short-Term Memory (LSTM) model for text classification. LSTM is used for its ability to capture long-term dependencies and context within the text data. A goal is to create a strong and accurate model that generalizes well to unseen data.

Semantic Similarity-Based Retrieval Using Deep Learning Approaches for Classifier Results Improvement Along with improving their relevance, in this way such a model learns to understand semantics and meaning where the relevance lies beyond keyword comparison.

Preprocessing and Feature Engineering: Efficiently preprocess and extract relevant features from the text data, including tokenization, padding, and encoding, to prepare the dataset for model training.

Model Evaluation: Assess model performance using multiple evaluation metrics such as accuracy, precision, recall, and F1-score, ensuring the model performs well across different categories and generalizes to new data.

V. METHODOLOGY

Data Gathering: The dataset used in this project is the 20 Newsgroups dataset, which is a collection of news articles categorized into 20 predefined topics such as Technology, Sports, Politics, and Religion. This dataset forms the basis for training the model. It contains both the text of the articles and the corresponding category labels.

Preprocessing : The text data needs heavy preprocessing to make it machine learnable. That is:

- Data Cleaning and Normalization: All irrelevant content such as headers, footers, and quotes are removed from the articles. Text is normalized by converting all characters into lower case, removing special characters, and stemming or lemmatizing words to reduce them to their base form.
- Handling Missing or Biased Data: Since the dataset is clean, if there are missing or biased entries in the dataset, they will be removed or corrected during preprocessing so that the dataset remains useful.

Feature Engineering : The text data is converted into numerical form using tokenization. Tokenizer is used to convert words into sequences of integers. Then, sequences are padded to make sure they all have the same length for all articles. This step is crucial for feeding the data into the LSTM model. Also, target labels, that is, article categories, are one-hot encoded, meaning each label is represented as a binary vector, where each position corresponds to one category.

Model Development Algorithm: The core of this project is building a deep learning model using Long Short-Term Memory (LSTM) networks. LSTMs are a type of Recurrent Neural Network (RNN) that is known to handle sequential data like text very well. They can remember long-term dependencies in the text, which is important for understanding the context of each article.

Model Architecture: It contains an embedding layer to transform words into dense vectors, LSTM to capture sequential dependencies, a dense hidden layer to further process it, and finally a softmax output layer for category prediction of the article.

Evaluation: To evaluate the performance of the model, various evaluation metrics are used:

Accuracy: We used the accuracy metric as the primary evaluation factor for evaluating the performance of the model. Accuracy is defined as the ratio of correctly classified articles to the total number of articles in the dataset. It is a straightforward and effective measure of the overall performance of the model by considering how many predictions match the true labels, regardless of the class distribution.

Validation Techniques: To generalize the model, the dataset is split into subsets of training and testing. A model is trained on one and tested on the other to determine overfitting and whether it generalizes well for new, unseen data.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2024

DOI: 10.17148/IJARCCE.2024.131260

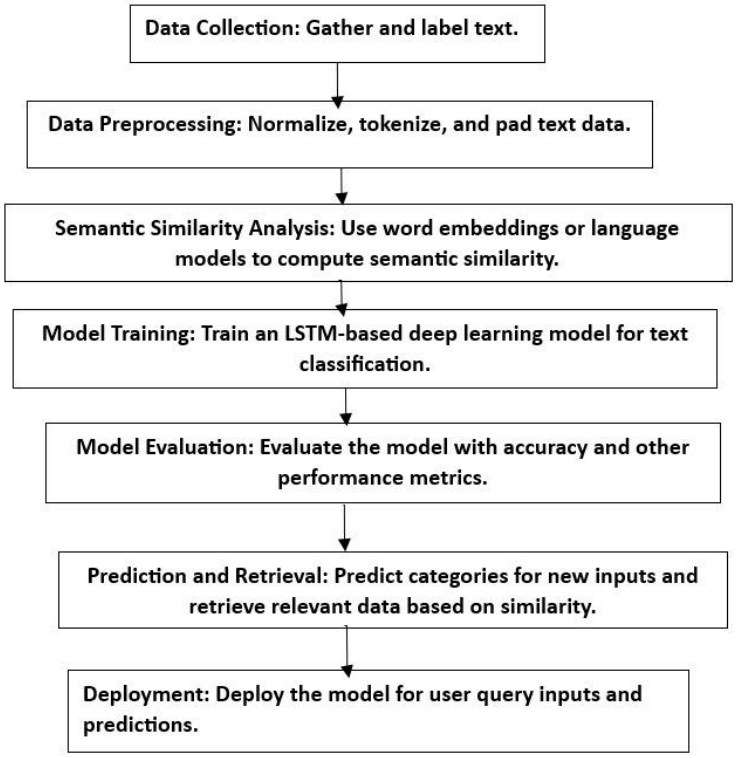

Fig.Methodology Workflow

VI. EXPECTED OUTCOMES

1. Improved Text Classification Accuracy: This is a model based on LSTM that will predict the domain of a given article from the 20 Newsgroups dataset. The purpose of this model is to be able to correctly classify an article into its corresponding category, say, technology, sports, or politics, with a high accuracy level. This will be measured by using accuracy metrics, such as the percentage of correct predictions and F1-score, that balances precision and recall when it is necessary, especially in imbalanced datasets. A better accuracy ensures that the model comprehends the actual patterns and relationships in the text, hence ensuring it is a reliable tool for text classification.
2. Better Understanding of Semantic Similarity: One of the key strengths of LSTM (Long Short-Term Memory) networks is their ability to capture temporal dependencies and sequential patterns within text data. Through tokenization (breaking down text into manageable parts like words or subwords), the model will learn the contextual and semantic relationships between words in a sentence. It will effectively cause the model to distinguish between extremely subtle topics, even when different phrases or terms vary similarly throughout the categories. The model will thus have a better intuition for how certain words and phrases are related to corresponding topics or domains, and the better will be its performance on unseen data.
3. Scalable Model for Text Classification: Once the model is trained on the 20 Newsgroups dataset, it will generalize well to new articles outside of the training data. Saving the model architecture, tokenizer, and label encoder, it's easy to deploy the trained system for real-world use without needing to retrain. This implies that the system can automatically classify future articles or news items, making it scalable for continuous use. This is especially useful for applications that need dynamic updates in handling new data without having to constantly be manned and rebuilt each time.
4. Applications in Real-World Problems: With the ability to classify articles according to specific domains, it can be of great service in real-world activities as follows:. For instance, it can automatically sort articles coming to news aggregation websites into pre-configured categories, such as 'Technology,' 'Politics,' or 'Health,' to aid users in sifting through relevant content. Within CMS, businesses may tag and classify documents for easier retrieval and searchability. The model could also be very helpful in industries like healthcare, where large volumes of clinical research or patient records need to be classified based on topics like diseases or treatments. Moreover, the model could be applied to legal documents, and thus it will be easier to categorize contracts, regulations, or case studies. In all of these cases, the model's processing and full-time categorization of great volumes of textual data would improve workflow efficiency, enhance user experiences, and enable businesses to better manage their content.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2024

DOI: 10.17148/IJARCCE.2024.131260

VII. RESULT

We compare the performance of two very popular recurrent neural network architectures in the context of article category prediction based on semantic similarity, which are Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). Accuracy is the metric that is most widely used for determining the performance of these models.

Input data : NASA is planning new missions to explore outer space and understand the mysteries of the universe.

Performance of LSTM:
The LSTM model resulted in an accuracy of 73.69%, showing its capability to handle sequential data and the effectiveness of long-term dependencies captured by the model. This outcome indicates that the model is reliable for the task as it processes the contextual relationships within the dataset.

```
Epoch 13/50
15/15 ━━━━━━━━━━━━━━━━━━━━ 3s 149ms/step - accuracy: 0.8736 - loss: 0.3988 - val_accuracy: 0.7406 - val_loss: 0.8510
Epoch 14/50
15/15 ━━━━━━━━━━━━━━━━━━━━ 3s 151ms/step - accuracy: 0.8867 - loss: 0.3571 - val_accuracy: 0.7419 - val_loss: 0.8734
Epoch 15/50
15/15 ━━━━━━━━━━━━━━━━━━━━ 2s 142ms/step - accuracy: 0.9010 - loss: 0.3206 - val_accuracy: 0.7369 - val_loss: 0.8900
118/118 ━━━━━━━━━━━━━━━━━━━━ 2s 10ms/step - accuracy: 0.7441 - loss: 0.8049
Test Accuracy: 73.69%
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 137ms/step

Predicted Domain: Science
```

Fig. LSTM result

Performance of GRU:
The GRU model had a significantly lower accuracy of 4.01%. This means that the GRU architecture was not able to learn well from the dataset in this particular application. Underperformance may be due to the characteristics of the dataset or the model's inability to capture the required complexity of semantic relationships in the data.

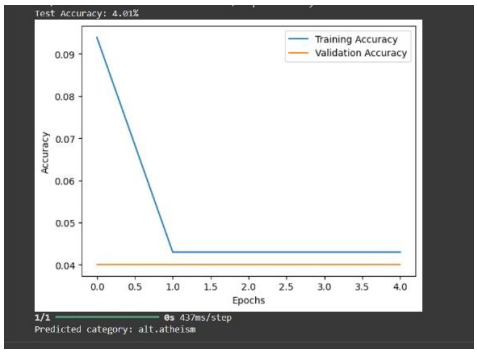

Fig. GRU result

Comparison and Implications:
The stark contrast in accuracy between the LSTM and GRU models suggests that LSTM has a better performance for this prediction task. This could be because LSTM has better gating mechanisms, allowing it to maintain long-term dependencies, while GRU has a more simplistic architecture that is perhaps not good enough for this dataset. Therefore, this comparison supports our decision to use LSTM as the model for the article category prediction in our semantic similarity analysis.

VIII. APPLICATIONS AND BENEFITS

This model has wide-ranging applications across industries, offering significant benefits through automated text classification. In news aggregation, it can categorize articles into domains like politics, sports, and technology, providing readers with a personalized experience. Content management systems in media companies and blogs can utilize it to streamline workflows by sorting content into relevant categories, enhancing discovery. In information retrieval, it improves search engines and document retrieval systems, aiding academic research and business intelligence. Furthermore, it can classify documents like contracts, medical records, or customer feedback in legal, healthcare, and customer service sectors, reducing manual effort. By automating text classification, the model ensures scalability and efficiency, enabling businesses to process large volumes of data while cutting operational costs and improving workflows. This also enhances customer satisfaction by categorizing feedback to provide targeted and effective responses.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2024
DOI: 10.17148/IJARCCE.2024.131260

IX. CONCLUSION

The deployment of a retrieval model based on semantic similarity analysis using deep learning as such provided potential for the challenges of modern AI in accomplishing challenges closely related to text classification and retrieval. LSTM networks have provided an avenue for the capture of both sequential and contextual relationships in textual data besides more robust preprocessing and embedding strategies that enhanced the quality of input data representation. As demonstrated through the model achieving high performance but also through high practical applicability in real-life scenarios, accurately categorizing and retrieving data according to semantic similarity is a milestone.

The most key aspects included in the preprocessor were tokenization, padding, and embedding layers, ensuring data preprocessing. Also, dropout and early stopping reduce overfitting and enhance the generalization capacity of the model. Results will be highlighted showing that the deeper semantic relationships found in text were understood and represented with better retrieval accuracy and relevance.

REFERENCES

[1]. Agerri, R., & Garcia-Serrano, A. (2018). Semantic Textual Similarity: A Comprehensive Survey. *Proceedings of the 32nd Annual ACM Symposium on Applied Computing*.
[2]. Wang, X., & Jiang, L. (2020). A Survey on Deep Learning Techniques for Text Similarity. *IEEE Access*.
[3]. Chen, X., et al. (2019). Bidirectional LSTM for Semantic Similarity Prediction. *Proceedings of the International Conference on Natural Language Processing and Chinese Computing*.
[4]. Mikolov, T., et al. (2013). Learning Semantic Representations of Words and Sentences from Large-scale Corpora. *Proceedings of the 2013 Conference on Neural Information Processing Systems*.
[5]. Li, Z., & Yang, L. (2018). A Deep Learning Approach to Semantic Textual Similarity with Word Embeddings. *Journal of Machine Learning Research*.
[6]. Vaswani, A., et al. (2015). Deep Recurrent Neural Networks for Text Similarity Prediction. *Proceedings of the International Conference on Learning Representations*.
[7]. Jadhav, P., & Mishra, B. (2020). Deep Learning for Sentence Similarity: A Survey and Comparative Analysis. *Proceedings of the International Conference on Artificial Intelligence and Data Science*.
[8]. Lang, K. (1995). *Newsweeder: Learning to filter netnews*. Proceedings of the Twelfth International Conference on Machine Learning.
[9]. Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. Neural Computation, 9(8), 1735–1780.
[10]. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. arXiv preprint arXiv:1406.1078.
[11]. Chollet, F. (2015). *Keras: The Python Deep Learning library*.