Analyzing Multilingual News Article Similarity

- Rutvik Shah
- Rohith Adhitya Chinnannan Rajkumar

Abstract

With the media increasingly becoming a crucial part of everyday life, publishers often compete to be the first to publish current affairs in articles. Given the increasing volume of controversial news, it is difficult to check the integrity of the news article being published. In this project, we will create a system to assess multilingual news articles' similarities to identify whether the two articles share the same story. This project is focused on evaluating the similarity of real-world affairs. In our project, a comprehensive dataset of multilingual news articles covering various topics, geographical regions, and languages has been used. The dataset used for this project is taken from the SemEval- 2022 Task which is focused on multilingual news articles. In our approach, the Sentence Bidirectional Encoder Representations from Transformers (SBERT) model is used as the encoder and transformer to encode the articles. The Natural Language Processing model that is employed is subjected to the use of sequential regression and the NLP (Natural Language Processing) layers will play a pivotal role in the project as they will be used for classification and text encoding to determine whether two articles are discussing the same event. A framework for performance evaluation of the system using metrics such as MAE (mean absolute error) score is created which evaluates the accuracy of the model. Using parameter tuning on the multi-layer model we have achieved a good score on our validation set.

I. INTRODUCTION

The task of evaluating multilingual news articles for real- world events is quite invaluable and comes with several complexities. These complexities involve not only linguistic diversity but also differences in cultural, political, and social contexts. News articles are full of contextual information that influences how a given event is presented. Analyzing news articles helps individuals become more educated by understanding how news is constructed, recognizing different journalistic styles, and differentiating between reliable and unreliable sources. Apart from this, news integrity is also a factor that needs to be seriously addressed. The media plays a vital role in society by holding those in power accountable. News integrity is necessary to maintain public trust in the media. Inaccurate or biased reporting can lead to misinformation, confusion, and social concerns [3]. This in turn gives rise to a more complex integrity issue, that is, multilingual news similarity and analysis. Multilingual news similarity is beneficial for cross-cultural understanding and decision-making on a global scale. It helps break down language barriers and promotes a harmonious global community in general. News in one language may have a different focus or interpretation than the same news in another language. Analyzing multilingual news similarity allows individuals to access multiple sources and viewpoints, enriching their understanding of events and reducing the risk of relying solely on one viewpoint. Multilingual news similarity helps identify such differences and allows users to cross-verify information to ensure accuracy.

This project was mainly inspired by the SemEval 2022 Task 8 [1] which addresses quantifying news similarity. This is a challenge particularly because when it comes to news articles, apart from the event that is being explained, there are various factors to take into consideration such as stories including descriptions, people, and entities that may appear in other dissimilar stories.

We used the dataset having a collection of 10,000 pairs of news articles published in 18 different languages. Considering dimensions, participants were told to estimate the overall similarity of news article pairs. The dimensions include geographic, temporal, and narrative similarity. By using a diverse and multilingual dataset, this data allowed participants to develop models that can effectively measure news story similarity, enabling clustering, identification of event coverage, and comparison of news outlets. The advantages lie in its practical application for structuring vast amounts of daily news articles, supporting research in media and communication studies, and fostering progress in automated methods for assessing news article similarity across different languages and dimensions. The motivation to work on this was to effectively be able to address the challenges faced in terms of controversial news, and integrity of news articles and to broaden readers' horizons and allow them to understand events from different perspectives.

Tasks such as text classification and entity recognition are achieved using various NLP models [2]. For text and document classification, NLP models, such as convolutional neural networks (CNNs), or transformer models (like BERT), are trained to classify text into predefined categories. Regression models analyze textual features to predict numeric outcomes such as engagement metrics or popularity scores. Its multifaceted capabilities ensure efficient multilingual communication through features like machine translation and therefore remove language barriers. In addition, NLP's proficiency in extracting relevant information from vast textual data is helpful for data analytics.

The reason to work with the Sentence-BERT model is that it is a highly powerful pre-trained language model that can significantly enhance the performance of our project while evaluating article similarity. The first step will be labeling the datasets to indicate whether the articles are matching. By training the model, the model learns to understand the similarities between two articles which it can later use to classify them. After this, it will produce dense embeddings for each article, which represent the content in a highly informative way. This allows us to compare articles based on their syntactic content rather than just keyword matching. The embeddings are then used to calculate the similarity between pairs of news articles. By breaking down language barriers and systematically assessing content similarities, this project has the potential to enhance the reliability of news sources. It gives users access to a broad area of viewpoints on the same global events, thereby contributing to a richer understanding of complex issues in today's information-driven world.

II. BACKGROUND

Gracia et al [4], proposed selective filtering on RSS news feeds using the Fuzzy-Set IR model. The content to be filtered was redundant news entries from all the news feeds. Several text similarity detection systems have been discussed in this paper such as commands in UNIX/LINUX, SIF, COPS, and SCAM but the main drawback of these approaches was they showed limitations in accurately identifying similarities between documents. The model uses correlation factors such as keyword connection, co-concurrence frequency, and distance which is the degree of similarity between two words. The authors experimented on 26 RSS news articles and manually counted 17 matching pairs. The results show key-connection and co- concurrence matrices to be around 50% accurate whereas the distance matrix is 94% accurate. Using the IR model and distance matrix allows for the automated removal of redundant and less informative RSS news articles. Therefore, facilitating the curation of more valuable and unique news articles for users.

The paper [5] focuses on improving stakeholder identification for managing an organization's institutional relationships. The identification process is done through the use of Named Entity Recognition (NER) models based on Bidirectional Encoder Representations from Transformers (BERT). After the process of vocabulary tokenization of both Multilingual and Portuguese models, 80% of the tokens were used for training and 20% for validation. The authors discuss the performance of three pre-trained masked models for their approach. These BERT models were compared to determine their performance in the Portuguese news datasets. The models were evaluated using the Micro F1 score after masking to find the best model. After the comparison of the scores of both datasets, it is observed that the Portuguese dataset had a score of 0.91 more than the multilingual dataset. The research concludes that the proposed NER models offer significant improvements over the existing rule-based approach for stakeholder identification.

This research [6] explores the application of the Sentence- BERT (SBERT) model in the setting of question retrieval within Community Question Answering (CQA) systems. While BERT has demonstrated good performance in sentence-pair regressions, SBERT is a more computationally efficient alternative by learning sentence representations from only one query question instead of requiring both questions to be given into the network. The study compares SBERT's performance to BERT4ECOMMERCE. It shows a slight decrease in accuracy but a significant reduction in the computational time for finding similar questions. The experiments involve fine-tuning SBERT with different loss functions demonstrating that the triplet loss function yields the highest Mean Average Precision (MAP) scores. After evaluating different pooling strategies, it is observed that CLS-token performs significantly better than the others. The paper suggests that SBERT has the potential for semantic similarity comparison, clustering, and information retrieval in CQA systems.

This paper [7] addresses the importance of text similarity measurement for information retrieval. The authors want to improve the accuracy of machine translation by fine-tuning Bidirectional Encoder Representation from Transformers (BERT) and training using the Double Siamese Text Convolutional Neural Networks (DSTCNN). The paper explores the development of textual similarity calculation such as the introduction of deep neural networks like AlexNet in image processing and advancements in word vector models such as word2vec. The authors conducted experiments using the STS-B dataset and achieved a Spearman's rank correlation coefficient of 88.1% for the two-stage model. This result gives an improvement of 1.6% over the first-stage model. The correlation coefficient suggests that the proposed two-stage model outperforms other algorithms. Therefore, it is observed that incorporating the Double Siamese Text Convolutional Neural Networks (DSTCNN) in the second stage enhances performance. Concluding the paper, the authors have highlighted that although Spearman's rank correlation coefficient of the STS-B mission has been improved, it still requires further study by researchers.

III. DATA COLLECTION

The data collected for the project is from the SemEval-2022 Task 8: Multilingual news article similarity [1] itself. The challenge includes a downloader that was used to obtain the data. The data consisted of a total of 10000 news article pairs having them in 18 different languages. The training data set contains 4,964 article pairs from multiple languages (English, German, Spanish, Arabic, Polish, Turkish, etc.) and gold standard similarity scores for six dimensions (Geography, Entities, Time, Narrative, Style, Tone), as well as the Overall score. The initial dataset in the directory is in HTML/JSON format. After importing the dataset, the Python global function was used to iteratively recover all the paths and file names. The global function is applied using Python's glob() function which is a module that is used to retrieve files and paths given a specific path argument. All the files containing various subparts of each news article were combined by parsing and generating Python dictionaries to store the titles, texts, and id-pairs. This was then stored in a single CSV file to be able to easily access the data for training.

The dataset consists of values that show the extent of similarity of the following attributes:

TABLE I. ATTRIBUTE INFORMATION

Attributes	**Information**
Pair_id	includes IDs for article pairs
Text1 and Text2	includes the content of the article
Title	Includes the title of the article pairs
Geography	includes the information about the geographic locations the article is focused on
Entities	includes the names of the entities involved in the news
Time	includes the time at which the news was about took place
Narrative	includes the type of narration of the article
Overall	includes the overall story of the article
Style	includes information about the writing styles of different articles
Tone	includes the tone of the news the article was published

The data collected is from January 2020 to June 2020. All the articles have more than a 100-word count. All the attributes are annotated using a Likert scale having 4 points namely “Very Dissimilar,” “Somewhat Dissimilar,” “Somewhat Similar,” and “Very Similar.” [1]. Among all the attributes, the key attributes that a primarily focused on are the pair_id to identify the article pairs, texts to analyze the content and ‘title’ to analyze the subject of the article pairs.

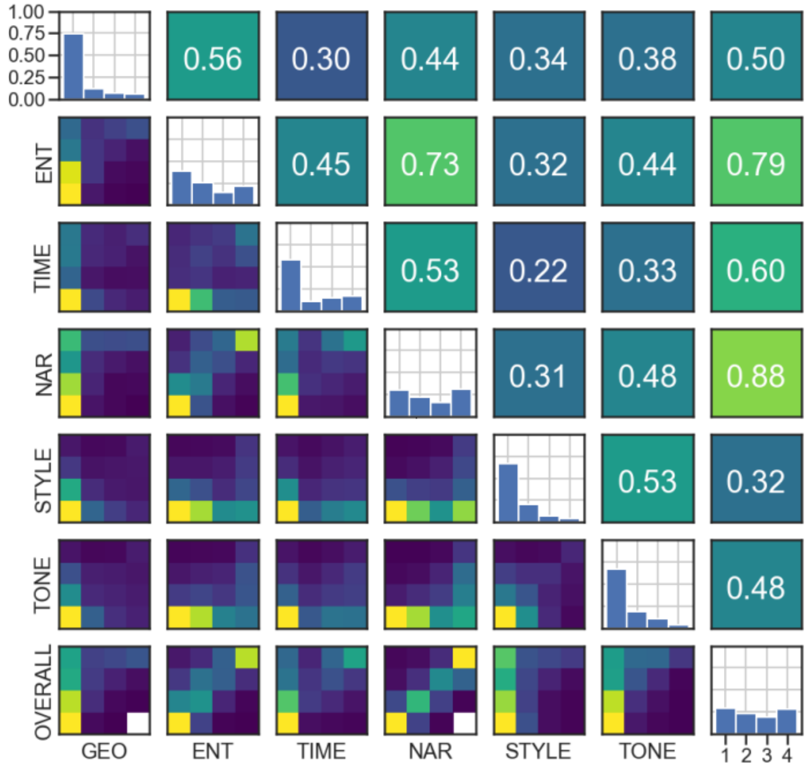


Fig. 1. Correlation between the attributes {Source: Image from [1]}

IV. DATA CLEANING AND PREPROCESSING

After data collection, the next step is to clean and pre- process the dataset to be further able to create a more reliable model having higher accuracy and minimal loss rate. The success rate for any NLP model depends highly on data quality, thus data cleaning is crucial. As mentioned earlier, the Python glob function is used to retrieve filenames and pathnames as the data extracted was stored in subfolders in HTML and JSON format files. These files are then converted and concatenated into one single CSV file. This CSV file containing all the labeled data was further used for training.

To avoid bias and inaccuracy in the model’s performance, the first step was removing all the missing values from the dataset. Handling null values can help increase the efficiency of the model and minimize faultiness. Any null values were dropped using the dropna() from the key attributes namely pair_id, title, and texts. These 3 attributes are the main features used to assess the similarity as they consist of the main idea of the news articles

To make the textual analysis more manageable and have a better representation of the data, all the punctuations, URLs, and numbers were removed using the ‘re’ module to manipulate strings.

(i) Removing Punctuations: The punctuations are not meaningful for text analysis so removing them will lessen the dimensionality of the data.

(ii) Removing URLs: URLs are considered noise in NLP models as they are irrelevant in terms of similarity analysis hence removing URLs will limit variability from the dataset.

(iii) Removing Numbers: In textual analysis, numeric elements have the least contribution so removing numbers gives the model a more simplified representation and supports the model to prioritize learning linguistic features.

After the comprehensive data cleaning and pre-processing, the dataset is now ready for robust analysis and modeling for deriving meaningful insights.

V. METHODOLOGY

A. NLP Model Preparation

For the sentence embeddings, the SBERT (Sentence Bidirectional Encoder Representations from Transformers) model has been used which is a pre-trained multilingual model that produces sentence embeddings [8] rather than converting each language into a particular language like English. The model ‘paraphrase-multilingual-mpvet-base-v2’ is a symmetric semantic search model having the capacity to encode more than 50 languages. This model was particularly chosen as it has performed the best among all the other pre-trained multilingual models.

The model was implemented to encode and transform each article into feature vectors. The semantic search is carried out by encoding each sentence into a 768-dimensional feature vector for subsequent analysis. Then each feature vector is stored in CSV files. This model is focused on fostering paraphrased sentences to identify similarities among various texts in different languages. As this model already has an encoder and transformer built-in, it is time-efficient and seems to be an appropriate choice to use when it comes to analyzing multilingual texts.

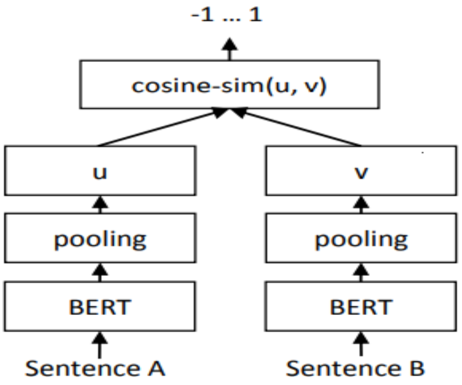


Fig. 2. SBERT architecture for regression function {Source: from [9]}

Further to load the pre-trained model, the TensorFlow and Keras function are used for model preparation. The regression model is developed with multiple hidden layers. The goal is to use a sequential regression model as it is known to be a good fit for NLP tasks such as natural language analysis and generation. Hence, a continuous input sequence of various articles to be processed is fed as an input and further, it is compared to learn the temporal and positional relationship between texts.

This model is created by adding linear layers such as a Dense layer to detect complex mapping among the input features, ReLU (Rectified Linear Unit) to reduce the possibility of negative values being labeled and thus constraining predictions, Dropout layer to generalize the model and prevent overfitting, and lastly, Batch normalization to transform and normalize the activation layers. For this dataset, the model is deployed with 7 dense layers, 2 dropout layers, 6 ReLU layers and 6 batch- normalization layers. With the successful preparation of the model, the train/test ratio is split into 70:30. After the split, 4700+ article pairs were there in the train set for further training and the rest for testing.

B. NLP Model training

With all the necessary imports the model is set to be trained. Further, all the needed variables were declared for the hidden layers along with deploying the linear layers onto the regression model. The model is fitted with initial batch size = 1 and epoch = 500. The model is compiled using Adam (Adaptive Moment Estimation) optimizer having a low learning rate which was implemented for precise convergence. A robust regression loss function - Huber loss was applied to prevent outliers. As well as a callback function that monitors the sets to early stopping as a regularization technique to stop after 5 epochs to avoid overfitting and save computational cost.

The model is trained over multiple epochs. Parameter tuning is done to the learning rate, dropout rate, and other hyperparameters to reach optimal results. Matplotlib is used to plot losses over training and validation sets to analyze the model learning on different epochs. Hence, experimenting with various hyperparameters to fine-tune the performance

With the initial training, the model having 500 epochs and a learning rate around 1e-7 or 1e-6 was observed to have overfitting on training data hence not a model that can be generalizable. To overcome this issue, the model was experimented with different hyperparameters until an optimal solution was achieved.

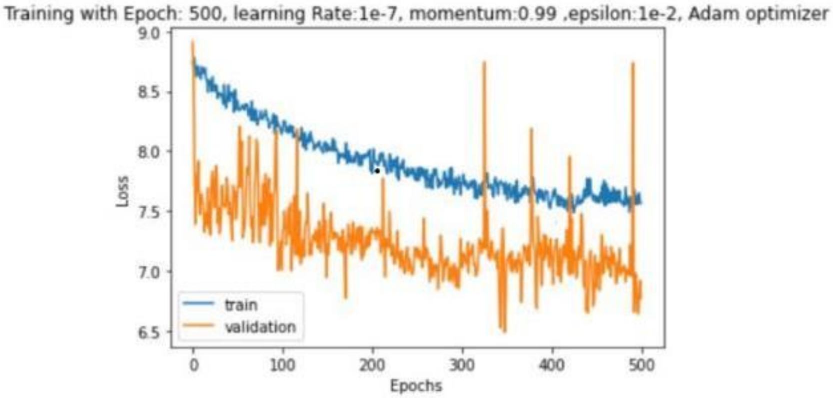


Fig. 3. Training model at epoch 500 and learning rate 1e-7

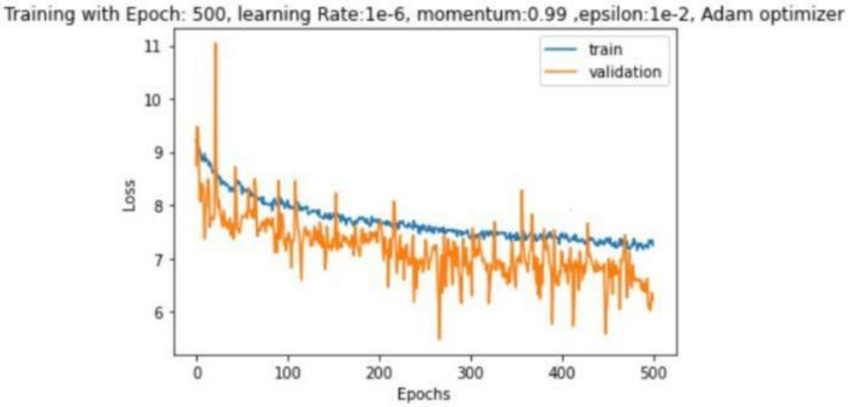


Fig. 4. Training model at epoch 500 and learning rate 1e-6

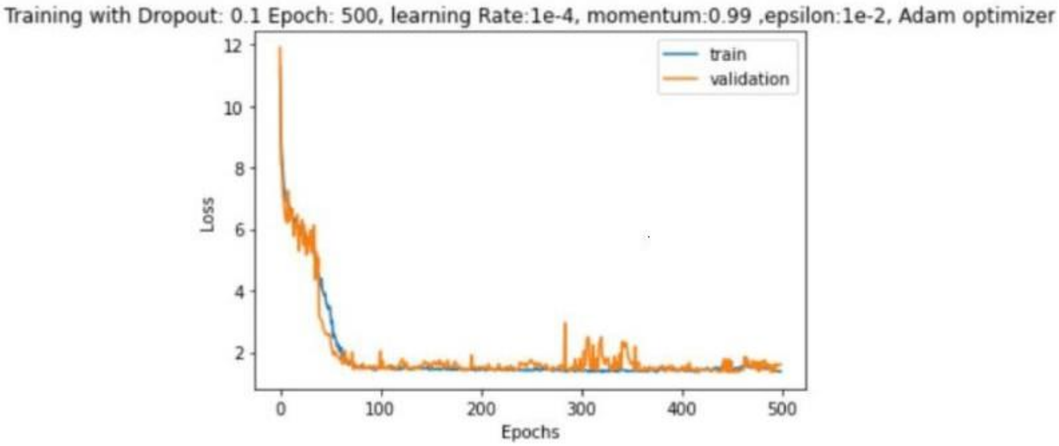


Fig. 5. Training model at epoch 500 and learning rate 1e-4

After successfully training the model multiple times, good training and validation scores were achieved. The above graph shows that both sets have a steady increase in accuracy over time which implies that the model is learning data well and is potentially generalizable. The model shows no overfitting.

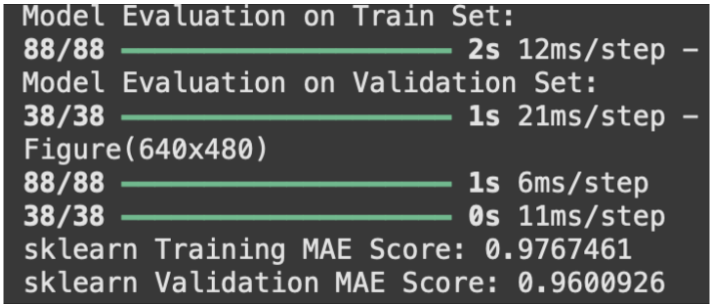
C. NLP Model Evaluation

The MAE score [10] is preferred to be used for sequential regression because of its reliable and interpretable nature. For the NLP model, it also has the efficiency to withstand outliers which is also as it divides equal weights around the errors. Hence, minimizing the overall loss in the model which is one of the primary goals of our project.

Following the training, the evaluation metric chosen is the MAE (Mean Absolute Error) score. Both training and validation sets are evaluated based on their respective MAE scores. The model after being fine-tuned resulted in the following prediction score:

On train set - 0.976

On validation set – 0.96



```
Model Evaluation on Train Set:
88/88 ━━━━━━━━━━━━━━━━━━━━ 2s 12ms/step -
Model Evaluation on Validation Set:
38/38 ━━━━━━━━━━━━━━━━━━━━ 1s 21ms/step -
Figure(640x480)
88/88 ━━━━━━━━━━━━━━━━━━━━ 1s 6ms/step
38/38 ━━━━━━━━━━━━━━━━━━━━ 0s 11ms/step
sklearn Training MAE Score: 0.9767461
sklearn Validation MAE Score: 0.9600926
```

Fig. 6. Screenshot of Model evaluation

VI. DISCUSSION AND LIMITATIONS

Nonetheless, after training the model multiple times with different hyperparameters, the model showed a good accuracy score. There were still numerous challenges faced to train the model. Some of the limitations are as follows:

(i) Complex model architecture: As it can be observed, with this dataset the model complexity was higher to understand each attribute. Also, resulting in the need to experiment using various hyperparameters.

(ii) Sensitivity to hyperparameter: It was noted that the graphs as well as the prediction score displayed vast differences when the parameters were changed. Hence, more amount of training time was needed.

(iii) Initial overfitting: Initially, the model showed overfitting but with parameter-tuning, we were able to reduce overfitting over time.

(iv) Performance dependency on data quality: The performance of the model largely depended on the quality of data. Hence, it can be said the data cleaning and pre-processing phase boosted the performance.

(v) Generalization issue to new data: There is a potential chance of the generalization issue. As mentioned, it will depend on the data quality. The model still needs more training to be able to work generally with other data.

For future scope, the model can be improved by adding more attributes and diverse set of metadata that includes images, URLs and more for precise results.

VII. CONCLUSION

Analyzing multilingual news articles is quite a handful task as it needs a lot of training to deploy an NLP-based model. The model is designed to have minimalistic loss and higher evaluation value. The training phase had some complexities but with parameter-tuning it was possible to achieve a good prediction score. Overall, leveraging NLP, this project represents a significant step towards using machine learning textual analysis in the context of news articles, with implications for understanding public opinion and enhancing decision-making processes.

VIII. REFERENCES

[1] X. Chen et al., ‘SemEval-2022 Task 8: Multilingual news article similarity’, in Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 1094–1106.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] V. V. Mayil and T. R. Jeyalakshmi, "Pretrained Sentence Embedding and Semantic Sentence Similarity Language Model for Text Classification in NLP," 2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP), VIJAYAWADA, India, 2023, pp. 1-5, doi: 10.1109/AISP57993.2023.10134937.

[3] A. Jalan, A. Gupta and P. Meel, "Comparing Results of Multiple Machine Learning Algorithms on a bilingual dataset for the Detection of Fraudulent News," 2023 12th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 2023, pp. 1-7, doi: 10.1109/MECO58584.2023.10154918.

[4] I. Garcia and Y.-K. Ng, "Eliminating Redundant and Less-Informative RSS News Articles Based on Word Similarity and a Fuzzy Equivalence Relation," 2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06), Arlington, VA, USA, 2006, pp. 465- 473, doi: 10.1109/ICTAI.2006.54.

[5] E. H. M. Da Silva, J. Laterza, M. P. P. Da Silva and M. Ladeira, "A proposal to identify stakeholders from news for the institutional relationship management activities of an institution based on Named Entity Recognition using BERT," 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Pasadena, CA, USA, 2021, pp. 1569-1575, doi: 10.1109/ICMLA52953.2021.00251.

[6] T. -T. Ha, V. -N. Nguyen, K. -H. Nguyen, K. -A. Nguyen and Q. -K. Than, "Utilizing SBERT For Finding Similar Questions in Community Question Answering," 2021 13th International Conference on Knowledge and Systems Engineering (KSE), Bangkok, Thailand, 2021, pp. 1-6, doi: 10.1109/KSE53942.2021.9648830.

[7] H. Zhengfang, I. K. D. Machica and B. Zhimin, "Textual Similarity Based on Double Siamese Text Convolutional Neural Networks and Using BERT for Pre-training Model," 2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 2022, pp. 107-111, doi: 10.1109/ICAIBD55127.2022.9820371.

[8] Z. Wei, X. Xu, C. Wang, Z. Liu, P. Xin and W. Zhang, "An Index Construction and Similarity Retrieval Method Based on Sentence- Bert," 2022 7th International Conference on Image, Vision and Computing (ICIVC), Xi’an, China, 2022, pp. 934-938, doi: 10.1109/ICIVC55077.2022.9886134.

[9] N. Reimers and I. Gurevych, ‘Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks’, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 08 2019, pp. 3473–3483.

[10] C. Willmott and K. Matsuura, ‘Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance’, Climate Research,

vol. 30, p. 79, 12 2005.

Note – Rutvik Shah carried out data collection, model training and hyperparameter tuning. Rohith Adhitya Chinnannan Rajkumar conducted performance evaluation, result analysis and both contributed in writing the paper. Contributions were equal.