Multilingual News Article Similarity based on BERT and RoBERTa

A Project Report Submitted in
Partial Fulfilment of the Requirements for the
8^{th} Semester B.Tech. Project

Submitted by

Samudranil Dutta	1912014
Rajdeep Paul	1912038
Veeranjaneyulu Chowdary Battula	1912103

Under the Supervision of

Dr. Partha Pakray
Assistant Professor
Department of Computer Science and Engineering
National Institute of Technology Silchar

Department of Computer Science and Engineering
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR

May, 2023

Department of Computer Science and Engineering
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR

Declaration

Thesis Title: **Multilingual News Article Similarity based on BERT and RoBERTa**

Degree for which the Thesis is submitted: **Bachelor of Technology**

We declare that the presented thesis represents largely our own ideas and work in our own words. Where others' ideas or words have been included, we have adequately cited and listed them in the reference materials. The thesis has been prepared without resorting to plagiarism. We have adhered to all principles of academic honesty and integrity. No falsified or fabricated data have been presented in the thesis. We understand that any violation of the above will causes disciplinary action by the Institute, including revoking the conferred degree if conferred, and can also evoke penal action from the sources which have not been properly cited or from whom proper permission has not been taken.

Signed: Somudranil Dutta, Battula Veeranjaneyulu Chowdary, Rajdeep Paul
16/5/2023. 16/5/2023 16/5/2023

Date: 16/5/2023

Department of Computer Science and Engineering
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR

It is certified that the work contained in this thesis entitled **Multilingual News Article Similarity based on BERT and RoBERTa** submitted by **Samudranil Dutta, Rajdeep Paul and Veeranjaneyulu Chowdary Battula** bearing Registration no **1912014, 1912038 and 1912103** respectively for the B.Tech. End Semester Project Examination May 2023 is absolutely based on their own work carried out under my supervision.

Place: NIT Silchar
Date: 16/05/23

Dr. Partha Pakray
Computer Science & Engineering
National Institute of Technology Silchar

Abstract

Numerous news articles are published daily in various languages, with over twenty crore news websites globally. According to reports, more than two million articles are published each day in various online news portals. Properly clustering news articles can have various benefits, such as recommending related articles and displaying similar content. Furthermore, identifying publications that cover the same story allows for cross-linguistic analysis of media consumption and attention. Narratives can differ in multiple ways, making it time-consuming to determine the similarity between two separate news articles. For example, two articles may share a considerable amount of text but they might discuss events that occurred years apart. This paper explores the use of deep learning-based models to establish the similarity between two different news articles.

Acknowledgements

We take this opportunity to express our sincere gratitude and heartily thanks to our supervisor Dr Partha Pakray, Department of Computer Science Engineering, National Institute of Technology Silchar for his continuous inspiration and valuable guidance at every stage of our research work. We would like to thank our Doctoral Committee Chairman and other members for their continuous evaluation and valuable constructive suggestions during this work. We would like to also thank all the faculty members of the Computer Science and Engineering Department of National Institute of Technology Silchar, for their administrative support during various phases of this work.

Your Name

Samudranil Dutta (1912014)

Battula Veeranjaneyulu Chowdary (1912103)

Rajdeep Paul (1912038)

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The increasing number of online news sources in multiple languages has created a need for effective methods to identify similar articles in different languages. This task is particularly challenging due to the diversity of language structures, vocabulary, and writing styles. In recent years, advancements in Natural Language Processing (NLP) have enabled the development of powerful language models that can handle multilingual text analysis. In this paper, we have explored the use of two advanced deep learning-based models, BERT and RoBERTa, for determining the similarity of multilingual news articles. The primary objective of this study is to evaluate the performance of these models in capturing the semantic and syntactic similarities between articles in different languages and to provide insights into their effectiveness for this task. The results of this research will have practical implications for the development of cross-lingual information retrieval systems and will contribute to the growing field of NLP for multilingual text analysis.

Natural language processing (NLP) is a branch of AI that enables computers to understand and respond to human language in text or speech form. NLP uses various techniques to analyze and interpret human language, such as speech recognition, part of speech tagging, word sense disambiguation, named entity recognition, co-reference resolution, and sentiment analysis. NLP combines rule-based and statistical methods to help computers perform tasks such as translation, voice control, summarization, and sentiment analysis. NLP is difficult because human language is not always clear and consistent. Computationally representing and analysing human languages is a challenging task. In recent times, many researchers have explored this domain. The field of NLP aims to make computers understand the statements and words in human languages. NLP can be categorized into two parts. Firstly, Natural Language Understanding (NLU). It involves extracting concepts, entities, emotions and keywords

from natural language that enable the machines to understand and analyse. Secondly, Natural Language Generation (NLG) creates meaningful internal representations of phrases, sentences, and paragraphs. NLP finds application in numerous areas such as machine translation, email spam detection, information extraction, summarization, question answering, text similarity, and more [11].

The most common way of measuring text similarity is to convert the text into vectors in vector space. The vectors are constructed using the words of the texts and then cosine similarity is used to calculate the similarity between the texts. The starting point for text similarity is to find the similarities between the words of the two texts. Words can exhibit similarity in both lexical and semantic aspects. Lexical similarity refers to words that have similar character sequences. Knowledge-based methods may face limitations when encountering informal words that are absent from their lexical database. Furthermore, methods relying on word embedding vectors may be subject to bias due to the nature of the corpus used to extract the values of these vectors [5].

Measuring the similarity between sentences and documents is a subject of research in natural language processing. One practical application is determining whether two news articles cover the same topic. The surge in online newspaper publications can be attributed to advancements in digital technology. In the modern age, where information is rapidly disseminated, readers must verify the authenticity of the news they are reading. False news and information can endanger individuals and even entire societies, making it crucial to verify sources and compare them with other news [3].

Tagging entities in text with their respective types is known as named entity recognition (NER). This task serves as an initial step for information extraction, where named entities in the text are identified and classified into predetermined categories such as names of individuals, organizations, locations, expressions of time, quantities, monetary values, percentages, and more. NER finds applications in various fields within natural language processing (NLP).

In this thesis, pre-trained BERT and RoBERTa have been used for calculating the similarity between two distinct news articles. The results obtained from both models along with the accuracy, precision, recall and F-score are reported. Also, a user interface is created to measure or calculate the similarity by directly pasting the two news articles.

CHAPTER 2

Literature survey

We reviewed some of the research in the field of multilingual news article similarity. While many studies covered foreign languages, we found that none of them covered Indian languages. Chen et al. [1] found that systems that utilised several elements of the article—including the headline, content, and publication date—as well as systems that fine-tuned or otherwise trained embeddings, outperformed those that didn't. Multilingual embeddings and translation were typically coupled in the best-performing systems. The optimal architectures, embedding models, or preprocessing to apply to the data, however, were not clearly agreed upon. There were obvious differences between languages, and additional effort is required to develop multilingual systems that function across various language pairings. When the news stories were dissimilar overall but had some resemblance in terms of their geographic focus, temporal focus, identified entities, and narratives, errors were more likely to occur.

Ritika et al. [3] carried out a comparison of three different techniques to gauge the semantic similarity between two news pieces on (almost) the same topic or event in two different languages (Hindi and English). The GoogleNews data sets were used to test the experiment. The three approaches are the Bag of Words Euclidean distance, the similarity of Cosine with tf-idf vectors, and the similarity of Jaccard with tf-idf vectors. All three of these techniques had encouraging results, but cosine similarity utilising tf-idf had the highest accuracy, recall, and F-measure scores (81.25%, 100%, and 76.92%).

Ishihara et al. [2] demonstrated that the Bi-Encoder design outperformed the Cross-Encoder. The former strategy produced better results, however, the latter frequently produces higher performance. CLS was the most effective pooling strategy in this task,

and the model with BERT-base-multilingual-uncased and BERT-base-multilingual-cased did the best. Also, it was noted that the performance declined as the maximum length shrunk. They discovered that multilingual models that had been pretrained performed better.

Goutam et al. [4] proposed work is focused on establishing an interpretable Semantic Textual Similarity (iSTS) technique for a pair of sentences which can explain why two sentences are entirely identical, only somewhat similar, or have some deviations. For three separate datasets, the effectiveness of the suggested method is confirmed. The assessed results demonstrate that the suggested strategy outperforms alternative iSTS methods. Most significantly, a Textual Entailment (TE) technique is created using the modules of the proposed iSTS approach. The entailment findings are found to greatly improve when chunk level, alignment, and sentence level features are integrated.

Hameed et al. [5] proposed a method for measuring the degree of semantic similarity between texts by fusing knowledge-based and corpus-based semantic data to create a semantic network that depicts the relationships between the compared texts and determines the degree of similarity between them. The GloVe pre-trained word embedding vectors were utilised as a corpus-based source, and the WordNet lexical database was employed as a knowledge-based source. Three separate datasets, the DSCS, SICK, and MOHLER datasets, were used to evaluate the suggested methodology. In terms of RMSE and MAE, a better results has been achieved.

Malte et al. [6] proposed a document similarity measure that takes into account the aspects of how two documents are similar. The proposed methodology was evaluated for research papers. A series of Transformer models such as RoBERTa, ELECTRA, XLNet, and BERT variations were applied and compared to an LSTM baseline. The experiments were performed on two newly constructed datasets of 172,073 research paper pairs from the ACL Anthology and CORD-19 corpus. SciBERT emerged as the best-performing system with F1-scores of up to 0.83.

Hung Chim and Xiaotie Deng [7] proposed a phrase-based document based on the Suffix Tree Document (STD) model similarity for computing the pairwise similarities of documents. The phrase-based document similarity is applied to the group-average Hierarchical Agglomerative Clustering (HAC) algorithm and develops a new document clustering approach. The experiments indicated that the new clustering approach was very effective in clustering the documents of two standard document benchmark corpora OHSUMED and RCV1. The results obtained from clustering outperformed the results of traditional single-word tf-idf similarity measures. It was concluded that the STD model can be considered as an expanded feature vector of the traditional single-word terms in the VSD model and thus the phrase-based document similarity works much better than the single-word tf-idf similarity measure.

Fabio et al. [8] proposed Context Semantic Analysis, which is a new knowledge-based method for computing inter-document similarity (CSA). A Semantic Context Vector, which can be taken from a knowledge base, is the foundation of this method for calculating inter-document similarity. According to experimental findings, CSA outperformed baselines produced on top of conventional approaches for the overall goal of inter-document similarity and attained performance comparable to those built on top of specialised knowledge bases. The Semantic Context Vector model was used to improve the results of cutting-edge technology that used comparable semantic enrichment for Information Retrieval tasks as well.

Xu et al. [9] proposed a Regression Model with Data Augmentation for Multilingual News Similarity. They developed a reasonably effective system for determining the similarity between a pair of news articles in multilingual and cross-lingual settings by utilising a variety of optimisation techniques, including data augmentation, head-tail combination, multi-label loss adapted R-Drop, and adding additional linear layers.

Heil et al. [10] proposed a model for evaluating the similarity between pairs of multilingual news articles. They used a Neural Machine Translation method to make it easier to compare multilingual articles with Sentence-BERT models. The suggested approach

performs consistently when estimating the similarity between monolingual and multilingual document pairs. The state-of-the art pretrained word and phrase embeddings produced a quick system with no computational overhead, enabling implementation without the need for graphics processing units.

Montalvo et al. [12] presented an approach for bilingual news clustering. The documents, the news, were represented only by means of cognate Named Entities (NE). They suggested a similarity metric for similarity based on a fuzzy rule framework. These rules attempted to take into account the significance of the category of identified entities in the news. Five different, equivalent news corpora in English and Spanish were used in the studies. Their method produced improved outcomes across all corpora, therefore it appeared ideal for bilingual news clustering.

Rupnik et al.[13] proposed a cross-lingual system for linking events in different languages. They also suggested a method for linking events and assessing the efficiency of various features. The final pipeline accurately links events while being scalable in terms of the volume of articles and the number of languages. According to the results of the trials, the core CCA-based features provide a solid basis that can benefit considerably from additional semantic-based features. There are two significant advantages to the technique, despite the fact that the addition of CCA-based features to semantic features did not result in significant improvement in speed. Initially, choosing a lesser number of potential clusters can speed up the linking process. Due to lack of linguistic resources, the method is robust to languages where semantic extraction is not possible.

In Semantic Textual Similarity, systems evaluate the level of semantic congruence between two text fragments. The creation of a unifying framework for merging various semantic components is one of its objectives. The paradigm that STS offers enables an extrinsic assessment of these modules. Additionally, such an STS framework itself might be assessed both internally and externally as a "grey box" or "black box" across a variety of NLP applications, including Machine Translation (MT), Summarization, Generation, Question Answering (QA), etc [15].

CHAPTER 3

Dataset Preparation

Preparation of the dataset involved collecting news articles from various news websites. Three datasets each of size 200, containing a pair of news articles are made. The first dataset consists of pairs of an English news article and a Bengali news article. The second dataset consists of pairs of an English news article and an Assamese news article. The third dataset consists of pairs of an English news article and a Telegu news article. The challenge was most of the regional news was not reported in the mainstream media. Similarly, regional newspapers do not publish the news that dominates the mainstream news media. The various websites from where English, Bengali, Assamese and Telegu news articles were collected are shown in Table 3.1, 3.2, 3.3 and 3.4 respectively.[1]

Manual tagging by human evaluators is done for each of the pairs. Each pair is assigned an **S** if the articles are similar and a **D** if dissimilar. Three human evaluators were employed for marking the news article pairs as similar or dissimilar as shown in Table 3.5, 3.6 and 3.7. If at least two of the three human evaluators marked it as similar, the news articles are considered similar. Else they were marked as dissimilar.

The first dataset contains an English news article and a Bengali news article. 122 pairs were marked as **S** and 78 were marked as **D**. The second dataset contains an English news article and an Assamese news article. 119 pairs were marked as **S** and 81 were marked as **D**. The third dataset contains an English news article and a Telegu news article. 120 pairs were marked as **S** and 80 were marked as **D**.

[1] https://github.com/Rajdeep-Paul-117/FYP-Dataset/tree/main/source

TABLE 3.1: English news articles source

Source	Number of articles
ThePrint	80
mint	70
India Today	60
The Economic Times	45
NDTV	90
The Times of India	75
Hindustan Times	65
CNN	35
Reuters	25
BBC	35
The Indian Express	20

TABLE 3.2: Bengali news articles source

Source	Number of articles
Anandabazar Patrika	40
ABP Ananda	40
Hindustan Times Bangla	10
News18 Bangla	10
Zee 24Ghanta	20
Tv9 Bangla	30
Sangbad Pratidin	20
oneindia Bengali	30

TABLE 3.3: Assamese news articles source

Source	Number of articles
Asomiya Pratidin	30
Niyomiya Barta	25
NorthEast now	25
Time8	25
DY365	25
News18 Assam	50
Gana Adhikar	20

TABLE 3.4: Telegu news articles source

Source	Number of articles
TV9 Telegu	50
Andhra Jyothi	50
Andhra Prabha	20
tv5news	40
Sakshi	15
News18 Telegu	25

TABLE 3.5: Human Evaluator statistics for English Bengali pair

Verdict	Evaluator 1	Evaluator 2	Evaluator 3
Similar	122	122	120
Dissimilar	78	78	80

TABLE 3.6: Human Evaluator statistics for English Assamese pair

Verdict	Evaluator 1	Evaluator 2	Evaluator 3
Similar	119	125	119
Dissimilar	81	75	81

TABLE 3.7: Human Evaluator statistics for English Telegu pair

Verdict	Evaluator 1	Evaluator 2	Evaluator 3
Similar	120	120	118
Dissimilar	80	80	82

TABLE 3.8: Dataset Summary

Language pair	Similar	Dissimilar
English Bengali	122	78
English-Assamese	119	81
English-Telegu	120	80

CHAPTER 4

Methodology

In this section, the methodology used has been discussed. The flowchart in Figure 4.1 provides an overall overview of the methodology. Each of the steps is explained below.

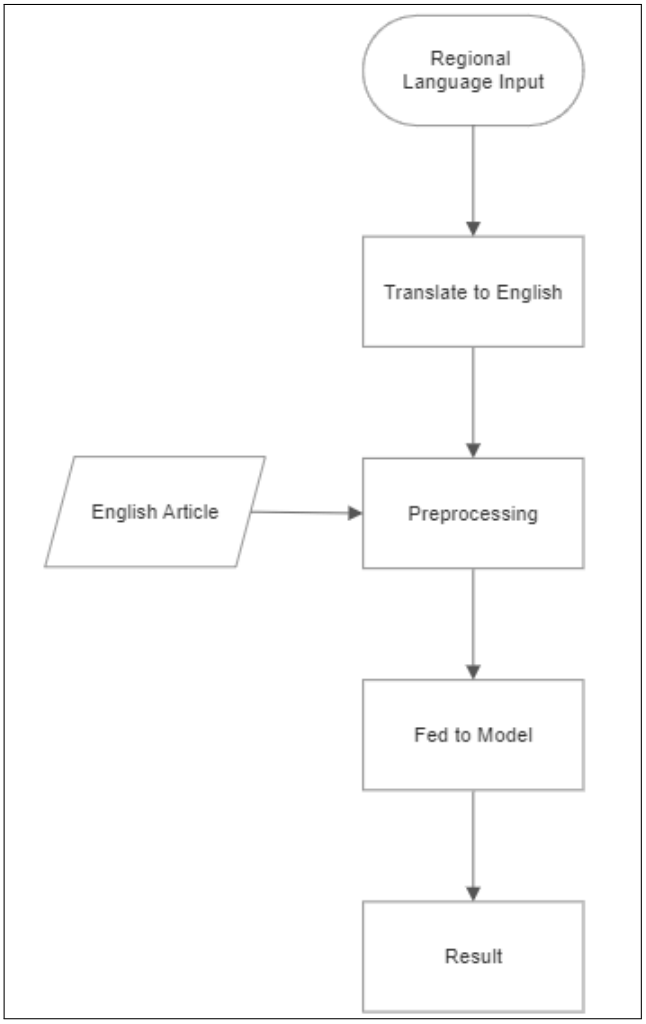

FIGURE 4.1: Flowchart

Translation: The regional language (Bengali and Assamese) is translated into English language using Microsoft Bing translator API. The API automatically identifies any language and translates it to the target language. As a preprocessing step, the stopwords were removed from both news articles, however, no significant improvement in results was observed. The steps are shown in the flowchart in Figure 4.1.

Preprocessing: Stopwords are removed from the articles and then they are fed to the models. Stopwords are words which occur frequently in texts and have very less significance.

Metric Used: In the case of Cosine similarity, the text is first converted to vectors and the similarity of two vectors is calculated by taking the dot product and dividing it by the magnitudes of each vector. Cosine similarity gives an output between 0 to 1. 1 means the documents are most similar and 0 means they are dissimilar to each other.

Models Used: BERT (Bidirectional Encoder Representations from Transformers) is a language model that leverages Transformer, an attention mechanism capable of learning contextual relationships between words or sub-words in a text. Traditional Transformer includes two mechanisms, an encoder that processes the text input and a decoder that generates task predictions. Unlike directional models that process text input sequentially, either from left-to-right or right-to-left, the Transformer encoder reads the entire sequence of words simultaneously, making it non-directional, although commonly referred to as bidirectional. As a result, the model can understand the context of a word by analyzing its surrounding context, both to the left and right.

The Robustly Optimized BERT Approach (RoBERTa) is a modified version of BERT that employs improved training methods, 1000% more data, and enhanced compute power. RoBERTa enhances the training process by eliminating the Next Sentence Prediction (NSP) task from BERT's pre-training and introducing dynamic masking, which changes the masked token during training epochs. Furthermore, larger batch-training sizes were found to be more effective in the training process.

Result: The result obtained is a percentage which actually tells us the extent of similarity between the two news articles and is compared to the threshold value. If

the result obtained is below the threshold value then the articles will be considered dissimilar. Else it will be considered similar.

CHAPTER 5

System Description

The collected news article pairs have to be passed through our system for determining whether they are similar or not. The front end is made using HTML, CSS and ReactJS. The backend is made in Flask, a web application framework written in Python (version 3.9). The GUI description is shown in Figure 5.1. The user will paste an English news article in the left textbox, and a news article in any regional language in the right textbox as shown in Figure 5.2. On clicking the translate button the regional language article gets translated into the English language as shown in Figure 5.3. Clicking on the extract named entities button will give a pop-up containing named entities like date, location, name, etc. in both articles as shown in Figure 5.4. Two more buttons for calculating similarity using BERT and roBERta respectively are added in the GUI. It will generate a pop-up showing the similarity in terms of percentage out of 100 as shown in Figure 5.5.

The navbar is present in the topmost part of the GUI containing different menus: History, Upload, Files, Dataset, Login, Signin and Logout. The history tab shows the user's previous article comparisons as shown in Figure 5.6. The upload tab shown in Figure 5.7 is for uploading files for calculating similarity scores directly. The files tab shown in Figure 5.8 is for viewing the previously uploaded files that can be downloaded as when needed. Figure 5.9 shows the user login and signup pages.

The screen in Figure 5.9 shows the pages for user registration and login.

FIGURE 5.1: Home Screen

FIGURE 5.2: Input Articles

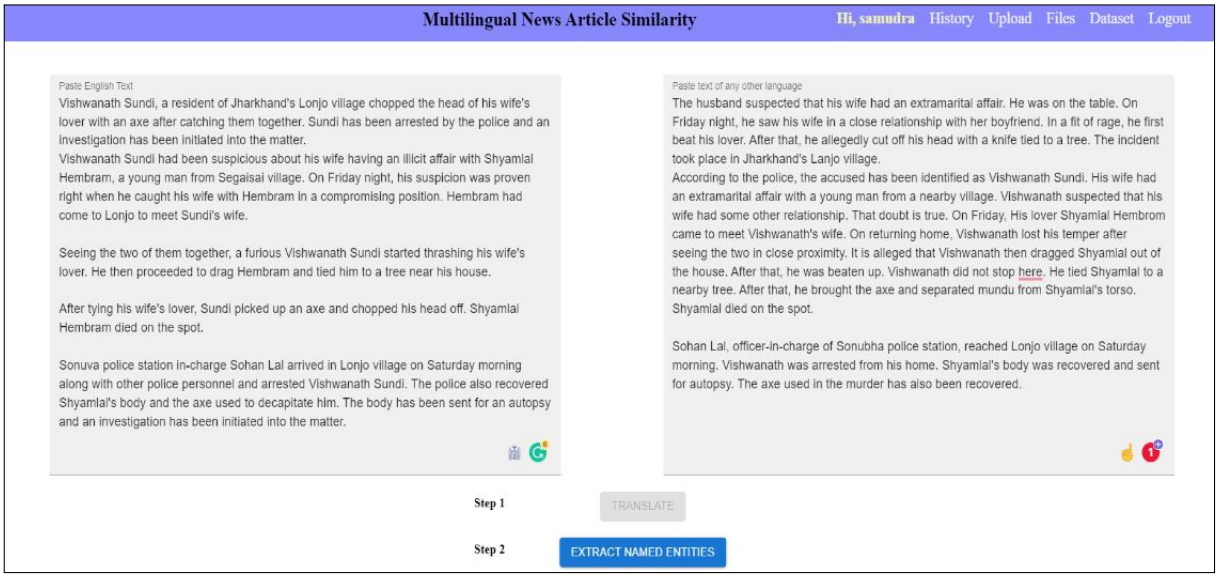

FIGURE 5.3: Translate

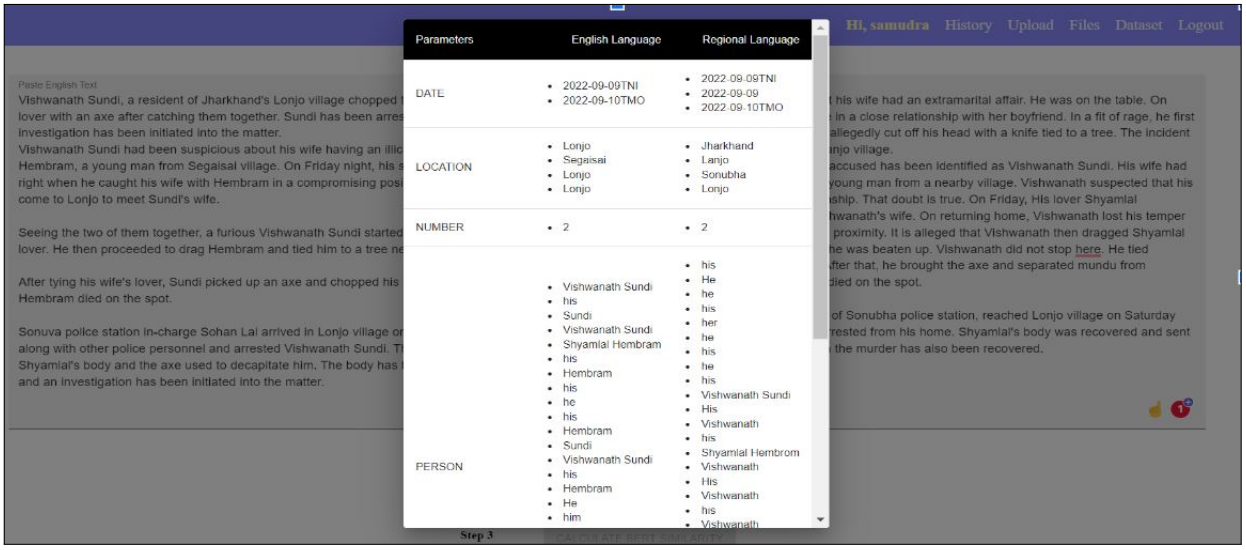

FIGURE 5.4: Named Entity Recognition

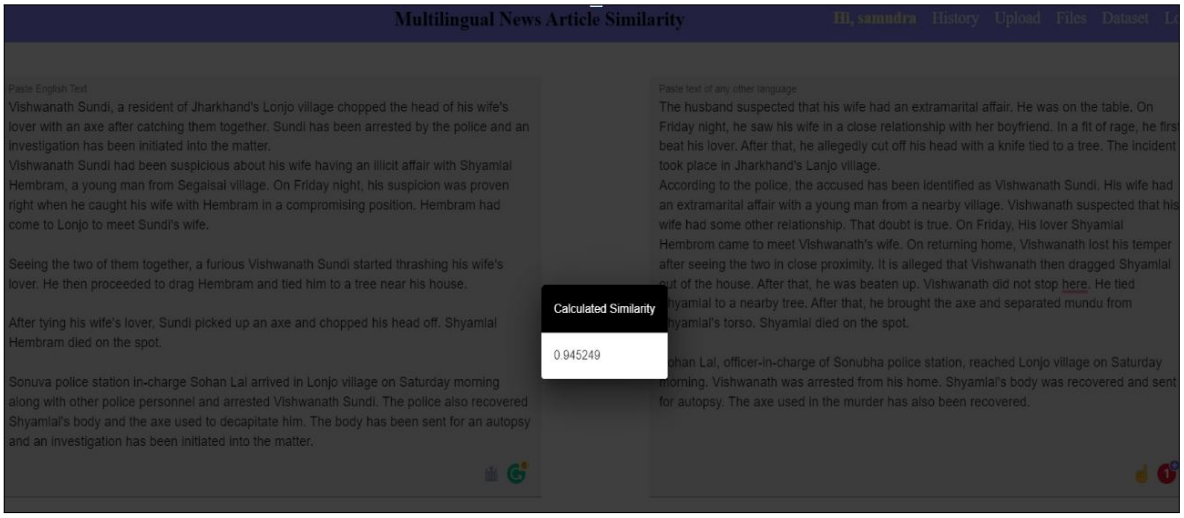

FIGURE 5.5: Result

FIGURE 5.6: History

FIGURE 5.7: File Upload

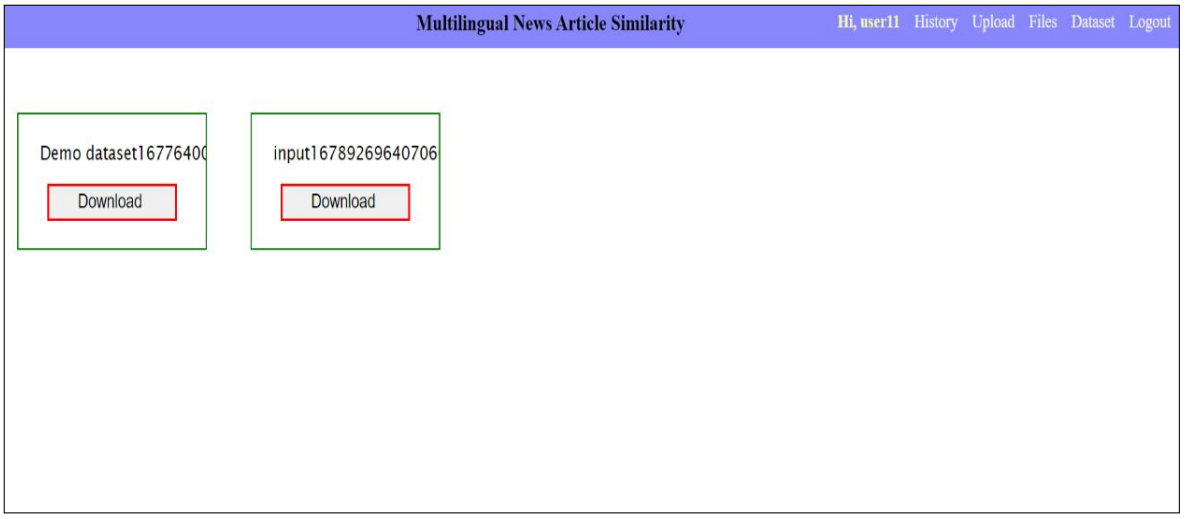

FIGURE 5.8: File Download

FIGURE 5.9: Login and Registration

CHAPTER 6

Experimental Results

We tested the BERT and RoBERTa models on all three datasets that we prepared. The result (percentage) we obtained was used to calculate the threshold value. If the result obtained is below the threshold value then the articles will be considered dissimilar. Else it will be considered similar. To calculate the threshold value, we plotted the accuracy vs threshold value graph. Keeping this threshold value as a reference, we evaluated the result obtained from testing our test data on our model. We also calculated the precision, recall and F-score.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{6.1}$$

$$PRECISION(P) = \frac{\text{TP}}{\text{TP+FP}} \tag{6.2}$$

$$RECALL(R) = \frac{\text{TP}}{\text{TP+FN}} \tag{6.3}$$

$$F - SCORE = \frac{\text{2PR}}{\text{P+R}} \tag{6.4}$$

TP:True Positive
FP: False Positive
TN:True Negative
FN: False Negative

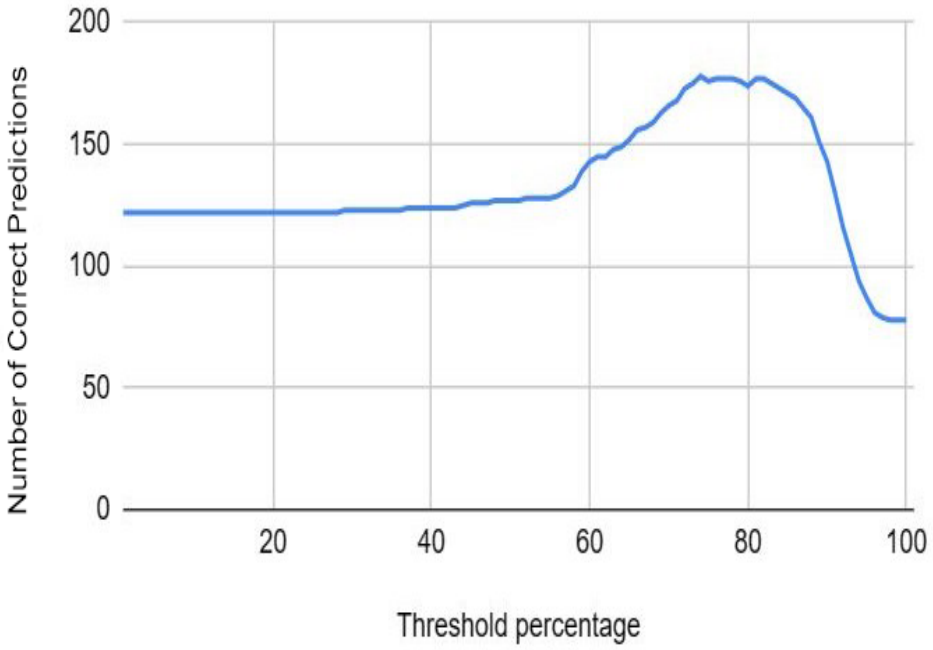

FIGURE 6.1: Number of correct predictions vs Threshold for BERT

6.1 BERT model on English-Bengali pair

The number of correct predictions vs threshold percentage for the BERT model on the English-Bengali pair is shown in Figure 6.1. From the graph, it can be seen that number of correct predictions is maximum (178) when the threshold is 74%. The accuracy, precision, recall and F-score are calculated from Table 6.1.

TABLE 6.1: BERT output statistics

BERT	**ACTUAL SIMILAR**	**ACTUAL DISSIMILAR**
SYSTEM TAGGED SIMILAR	119 (TP)	19 (FP)
SYSTEM TAGGED DISSIMILAR	3 (FN)	59 (TN)

$$ACCURACY = \frac{\text{Correct Prediction}}{\text{Total Predictions}} = \frac{178}{200} = 0.89 \tag{6.5}$$

$$PRECISION(P) = \frac{\text{TP}}{\text{TP+FP}} = \frac{119}{119+19} = 0.862 \tag{6.6}$$

$$RECALL(R) = \frac{\text{TP}}{\text{TP+FN}} = \frac{119}{119+3} = 0.974 \tag{6.7}$$

$$F - SCORE = \frac{\text{2PR}}{\text{P+R}} = 0.914 \tag{6.8}$$

6.2 RoBERTa model on English-Bengali pair

The number of correct predictions vs threshold percentage for the RoBERTa model on the English-Bengali pair is shown in Figure 6.2. From the graph, it can be seen that number of correct predictions is maximum (190) when the threshold is 56%. The accuracy, precision, recall and F-score are calculated from Table 6.2.

Table 6.2: RoBERTa output statistics

RoBERTa	**ACTUAL SIMILAR**	**ACTUAL DISSIMILAR**
SYSTEM TAGGED SIMILAR	117 (TP)	5 (FP)
SYSTEM TAGGED DISSIMILAR	5 (FN)	73 (TN)

$$ACCURACY = \frac{\text{Correct Prediction}}{\text{Total Predictions}} = \frac{190}{200} = 0.95 \tag{6.9}$$

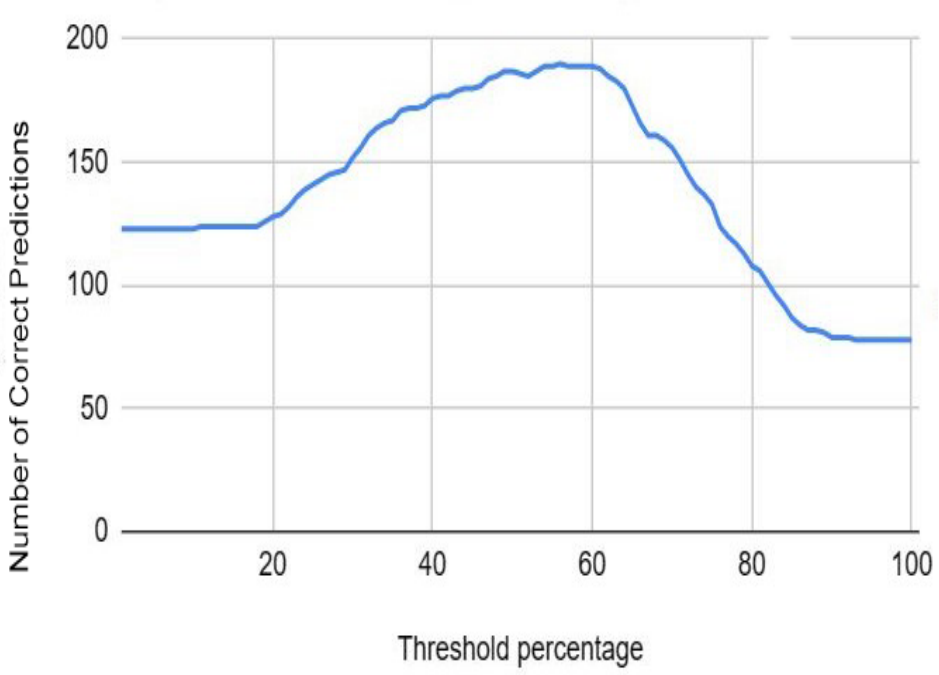

FIGURE 6.2: Number of correct predictions vs Threshold for RoBERTa

$$PRECISION(P) = \frac{\text{TP}}{\text{TP+FP}} = \frac{117}{117+5} = 0.959 \tag{6.10}$$

$$RECALL(R) = \frac{\text{TP}}{\text{TP+FN}} = \frac{117}{117+5} = 0.959 \tag{6.11}$$

$$F - SCORE = \frac{2\text{PR}}{\text{P+R}} = 0.959 \tag{6.12}$$

6.3 BERT model on English-Assamese pair

The number of correct predictions vs threshold percentage for the BERT model on the English-Assamese pair is shown in Figure 6.3. From the graph, it can be seen that number of correct predictions is maximum (193) when the threshold is 78%. The accuracy, precision, recall and F-score are calculated from Table 6.3.

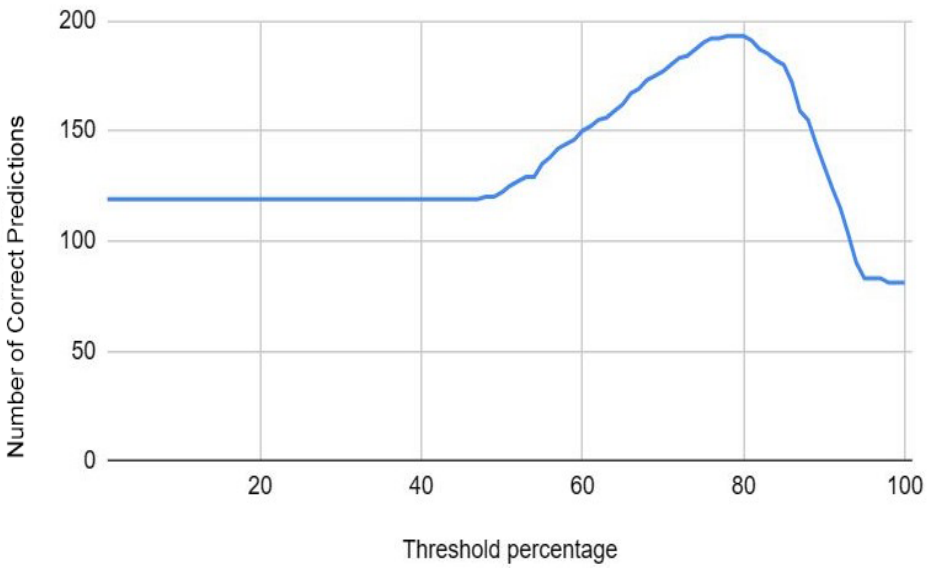

FIGURE 6.3: Number of correct predictions vs Threshold for BERT

TABLE 6.3: BERT output statistics

BERT	**ACTUAL SIMILAR**	**ACTUAL DISSIMILAR**
SYSTEM TAGGED SIMILAR	116 (TP)	4 (FP)
SYSTEM TAGGED DISSIMILAR	3 (FN)	77 (TN)

$$ACCURACY = \frac{\text{Correct Prediction}}{\text{Total Predictions}} = \frac{193}{200} = 0.965 \tag{6.13}$$

$$PRECISION(P) = \frac{\text{TP}}{\text{TP+FP}} = \frac{116}{116+4} = 0.966 \tag{6.14}$$

$$RECALL(R) = \frac{\text{TP}}{\text{TP+FN}} = \frac{116}{116+3} = 0.974 \tag{6.15}$$

$$F - SCORE = \frac{\text{2PR}}{\text{P+R}} = 0.97 \tag{6.16}$$

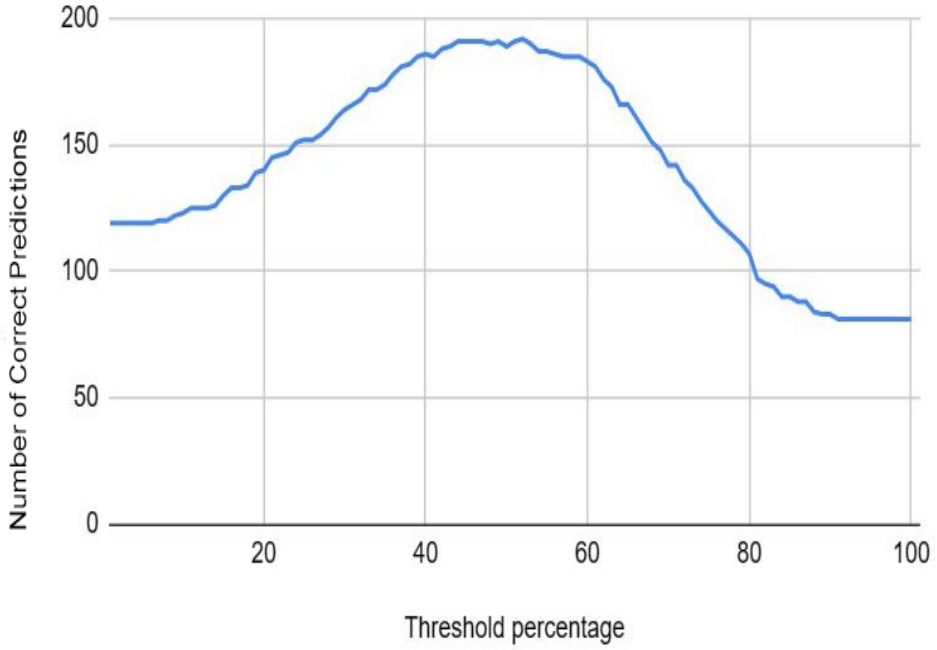

FIGURE 6.4: Number of correct predictions vs Threshold for RoBERTa

6.4 RoBERTa model on English-Assamese pair

The number of correct predictions vs threshold percentage for the RoBERTa model on the English-Assamese pair is shown in Figure 6.4. From the graph, it can be seen that number of correct predictions is maximum (192) when the threshold is 52%. The accuracy, precision, recall and F-score are calculated from Table 6.4.

TABLE 6.4: RoBERTa output statistics

RoBERTa	**ACTUAL SIMILAR**	**ACTUAL DISSIMILAR**
SYSTEM TAGGED SIMILAR	113 (TP)	2 (FP)
SYSTEM TAGGED DISSIMILAR	6 (FN)	79 (TN)

$$ACCURACY = \frac{\text{Correct Prediction}}{\text{Total Predictions}} = \frac{192}{200} = 0.96 \tag{6.17}$$

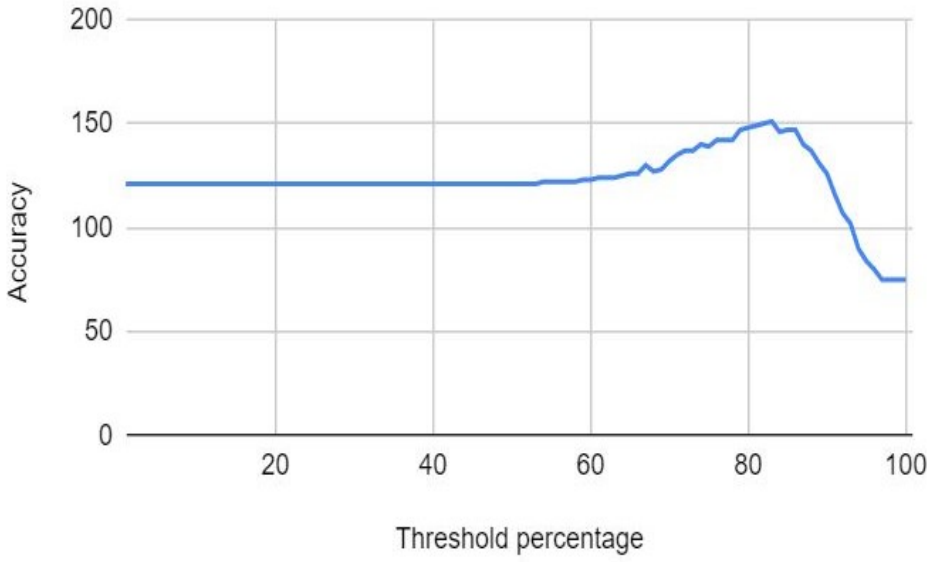

FIGURE 6.5: Number of correct predictions vs Threshold for BERT

$$PRECISION(P) = \frac{\text{TP}}{\text{TP+FP}} = \frac{113}{113+2} = 0.982 \tag{6.18}$$

$$RECALL(R) = \frac{\text{TP}}{\text{TP+FN}} = \frac{113}{113+6} = 0.949 \tag{6.19}$$

$$F - SCORE = \frac{2\text{PR}}{\text{P+R}} = 0.965 \tag{6.20}$$

6.5 BERT model on English-Telegu pair

The number of correct predictions vs threshold percentage for the BERT model on the English-Telegu pair is shown in Figure 6.5. From the graph, it can be seen that number of correct predictions is maximum (155) when the threshold is 83%. The accuracy, precision, recall and F-score are calculated from Table 6.5.

TABLE 6.5: BERT output statistics

BERT	ACTUAL SIMILAR	ACTUAL DISSIMILAR
SYSTEM TAGGED SIMILAR	99 (TP)	20 (FP)
SYSTEM TAGGED DISSIMILAR	25 (FN)	56 (TN)

$$ACCURACY = \frac{\text{Correct Prediction}}{\text{Total Predictions}} = \frac{155}{200} = 0.775 \tag{6.21}$$

$$PRECISION(P) = \frac{\text{TP}}{\text{TP+FP}} = \frac{99}{99+20} = 0.831 \tag{6.22}$$

$$RECALL(R) = \frac{\text{TP}}{\text{TP+FN}} = \frac{99}{99+25} = 0.798 \tag{6.23}$$

$$F - SCORE = \frac{\text{2PR}}{\text{P+R}} = 0.813 \tag{6.24}$$

6.6 RoBERTa model on English-Telegu pair

The number of correct predictions vs threshold percentage for the RoBERTa model on the English-Telegu pair is shown in Figure 6.6. From the graph, it can be seen that number of correct predictions is maximum (158) when the threshold is 57%. The accuracy, precision, recall and F-score are calculated from Table 6.6.

$$ACCURACY = \frac{\text{Correct Prediction}}{\text{Total Predictions}} = \frac{158}{200} = 0.79 \tag{6.25}$$

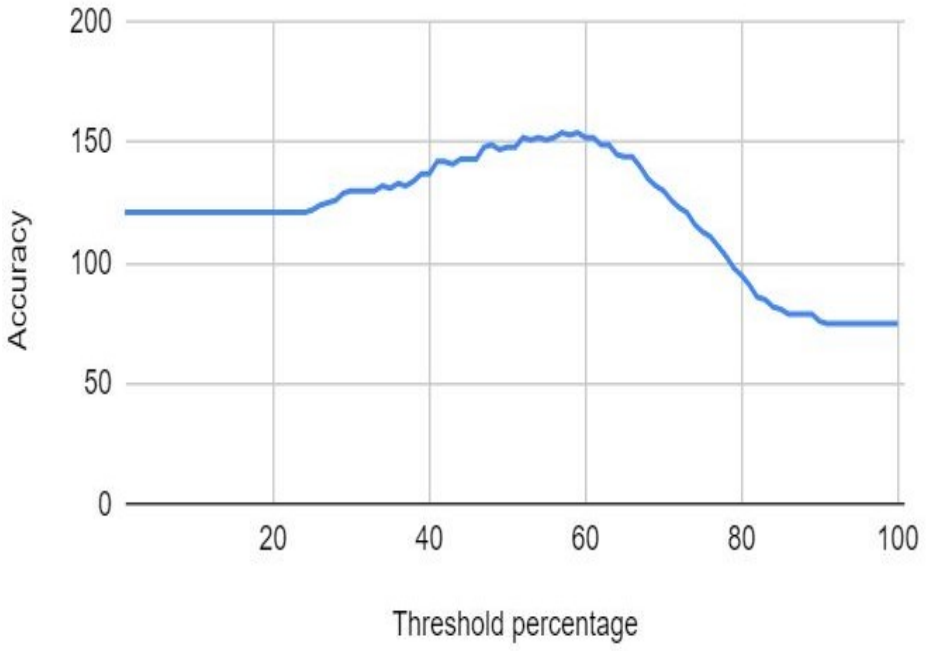

FIGURE 6.6: Number of correct predictions vs Threshold for RoBERTa

TABLE 6.6: RoBERTa output statistics

RoBERTa	**ACTUAL SIMILAR**	**ACTUAL DISSIMILAR**
SYSTEM TAGGED SIMILAR	91 (TP)	12 (FP)
SYSTEM TAGGED DISSIMILAR	30 (FN)	67 (TN)

$$PRECISION(P) = \frac{\text{TP}}{\text{TP+FP}} = \frac{91}{91+12} = 0.883 \tag{6.26}$$

$$RECALL(R) = \frac{\text{TP}}{\text{TP+FN}} = \frac{91}{91+30} = 0.752 \tag{6.27}$$

$$F - SCORE = \frac{2\text{PR}}{\text{P+R}} = 0.812 \tag{6.28}$$

CHAPTER 7

Discussions

The threshold value obtained for each language pair along with the model used is shown in Table 7.1. The results obtained from the dataset showed that both the BERT and

TABLE 7.1: Threshold value

Model	Language Pair	Threshold (in %)
BERT	English-Bengali	74
RoBERTa	English-Bengali	56
BERT	English-Assamese	78
RoBERTa	English-Assamese	52
BERT	English-Telegu	83
RoBERTa	English-Telegu	57

RoBERTa model tends to give the correct result for most of the cases. Still, there were some news article pairs on which the model did not perform well which is probably due to the complicated lexical variation and grammar of the articles. For English-Bengali pairs, the RoBERTa model performs much better than the BERT model which can be seen from the F-scores. However, for English-Assamese pairs, there is not a significant difference in the F-scores of both models. The accuracy, precision, recall and F-score for English-Telegu datasets for both models were quite lower when compared to the other two language pairs.
[1]

Some of the possible reasons causing these errors include differences in the news article length and translation-related problems. If the difference between the lengths of the two articles is significant, then there might be a situation where one news article has described the event in greater detail while the other news article hasn't covered it so extensively. Secondly, since we are using Microsoft Bing API to translate the news

[1] https://github.com/Rajdeep-Paul-117/FYP-Dataset/tree/main/error

TABLE 7.2: Error in English-Bengali pairs

Model	English	Bengali	Similarity %	Type of Error
BERT	News1	News2	71	FN
BERT	News3	News4	85	FP
RoBERTa	News5	News6	60	FP
RoBERTa	News7	News8	50	FN

TABLE 7.3: Error in English-Assamese pairs

Model	English	Assamese	Similarity %	Type of Error
BERT	News9	News10	80	FP
BERT	News11	News12	73	FN
RoBERTa	News13	News14	57	FP
RoBERTa	News15	News16	48	FN

TABLE 7.4: Error in English-Telegu pairs

Model	English	Assamese	Similarity %	Type of Error
BERT	News17	News18	85	FP
BERT	News19	News20	72	FN
RoBERTa	News21	News22	70	FP
RoBERTa	News23	News24	40	FN

articles from the regional language to the English language, some of this information might get lost while translating.

CHAPTER 8

Conclusion and Future Work

As reported in the previous section, our experiment's precision, recall and F-scores are impressive and accurate for English-Bengali and English-Assamese pairs. In future, we shall collect more pairs of news articles and test them on our model. Accordingly, the threshold might get adjusted. Besides this, we also plan to work on event detection.

In conclusion, this thesis evaluated the performance of BERT and RoBERTa models on three different language pairs English-Bengali, English-Assamese and English-Telegu for detecting similarities between news articles. The threshold value was calculated using the accuracy vs threshold value graph and used as a reference to evaluate the test data. The precision, recall and F-score were calculated for both models. The RoBERTa model showed better performance for English-Bengali pairs, whereas there was no significant difference between the two models for English-Assamese pairs. However, both models showed errors in some cases due to factors such as differences in article length and translation-related problems. Overall, the study demonstrates the potential of using pre-trained language models for text similarity detection and highlights areas for improvement in future research.

Bibliography

[1] X. Chen, A. Zeynali, C. Camargo, F. Flöck, D. Gaffney, P. Grabowicz, S. Hale, D. Jurgens, and M. Samory. SemEval-2022 task 8: Multilingual news article similarity. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1094–1106, Seattle, United States, July 2022. Association for Computational Linguistics.

[2] S. Ishihara and H. Shirai. Nikkei at semeval-2022 task 8: Exploring BERTbased bi-encoder approach for pairwise multilingual news article similarity. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1208–1214, 2022.

[3] R. Singh and S. Singh. Text similarity measures in news articles by vector space model using nlp. Journal of The Institution of Engineers (India): Series B, 102(2):329–338, 2021. https://doi.org/10.1007/s40031-020-00501-5

[4] Majumder, G., Pakray, P., Das, R. et al. Interpretable semantic textual similarity of sentences using alignment of chunks with classification and regression. Appl Intell 51, 7322–7349 (2021). https://doi.org/10.1007/s10489-020-02144-x

[5] Hameed, Naamah Hussien, Adel M. Alimi, and Ahmed T. Sadiq. "Short Text Semantic Similarity Measurement Approach Based on Semantic Network." Baghdad Science Journal 19.6 (Suppl.) (2022): 1581-1581.

[6] Ostendorff, M., Ruas, T., Blume, T., Gipp, B., Rehm, G. (2020). Aspect-based document similarity for research papers. arXiv preprint arXiv:2010.06395.

[7] H. Chim and X. Deng, "Efficient Phrase-Based Document Similarity for Clustering," in IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 9, pp. 1217-1229, Sept. 2008, doi: 10.1109/TKDE.2008.50.

[8] Fabio Benedetti, Domenico Beneventano, Sonia Bergamaschi, Giovanni Simonini. "Computing inter-document similarity with context semantic analysis." Information Systems 80 (2019): 136-147. https://doi.org/10.1016/j.is.2018.02.009

[9] Xu, Zihang, Ziqing Yang, Yiming Cui, and Zhigang Chen. "HFL at SemEval-2022 Task 8: A Linguistics-inspired Regression Model with Data Augmentation for Multilingual News Similarity." arXiv preprint arXiv:2204.04844 (2022).

[10] Heil S, Kopp K, Zehe A, Kobs K, Hotho A. LSX_team5 at SemEval-2022 Task 8: Multilingual News Article Similarity Assessment based on Word-and Sentence Mover's Distance. InProceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) 2022 Jul (pp. 1190-1195).

[11] Khurana, Diksha, Aditya Koli, Kiran Khatter, and Sukhdev Singh. "Natural language processing: State of the art, current trends and challenges." Multimedia tools and applications (2022): 1-32.

[12] Montalvo, S., Martínez, R., Casillas, A., Fresno, V. (2007). Bilingual News Clustering Using Named Entities and Fuzzy Similarity. In: Matoušek, V., Mautner, P. (eds) Text, Speech and Dialogue. TSD 2007. Lecture Notes in Computer Science(), vol 4629. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74628-7_16.

[13] Rupnik, Jan, Andrej Muhic, Gregor Leban, Primoz Skraba, Blaz Fortuna, and Marko Grobelnik. "News across languages-cross-lingual document similarity and event tracking." Journal of Artificial Intelligence Research 55 (2016): 283-316. https://doi.org/10.1613/jair.4780

[14] Agirre, E., Banea, C., Cardie, C., Cer, D.M., Diab, M.T., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G. and Wiebe, J., 2014, August. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In SemEval@ COLING (pp. 81-91).

[15] Chen, Zhongan, Weiwei Chen, Yunlong Sun, Hongqing Xu, Shuzhe Zhou, Bohan Chen, Cheng-Jie Sun, and Yuanchao Liu. "ITNLP2022 at SemEval-2022 Task 8: Pre-trained Model with Data Augmentation and Voting for Multilingual News Similarity." In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pp. 1184-1189. 2022.

[16] https://github.com/Rajdeep-Paul-117/FYP-Dataset/tree/main/source

[17] https://www.microsoft.com/en-us/translator/business/translator-api/

Multilingual News Article Similarity based on BERT and RoBERTa

ORIGINALITY REPORT

22%	17%	12%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to National Institute of Technology, Silchar Student Paper	5%
2	www.researchgate.net Internet Source	3%
3	aclanthology.org Internet Source	1%
4	www.computer.org Internet Source	1%
5	Davronov Rifqat Rahimovich, Safarov Ro'zmat Abdiqayum o'g'li, Abdumalikov Shoxrux Qaxramon o'g'li. "Predicting the activity and properties of chemicals based on RoBERTa", 2021 International Conference on Information Science and Communications Technologies (ICISCT), 2021 Publication	1%
6	deepai.org Internet Source	1%

7	Lecture Notes in Computer Science, 2013. Publication	1 %
8	ailab.ijs.si Internet Source	1 %
9	"Computational Linguistics and Intelligent Text Processing", Springer Science and Business Media LLC, 2018 Publication	1 %
10	link.springer.com Internet Source	1 %
11	Lecture Notes in Computer Science, 2007. Publication	1 %
12	Submitted to University of Technology Student Paper	1 %
13	Submitted to Mahidol University Student Paper	<1 %
14	"Proceedings of the Future Technologies Conference (FTC) 2020, Volume 1", Springer Science and Business Media LLC, 2021 Publication	<1 %
15	Submitted to Birla Institute of Technology and Science Pilani Student Paper	<1 %
16	Submitted to Universidad de Jaén Student Paper	<1 %

17	cs.nits.ac.in Internet Source	<1%
18	www.giovannisimonini.com Internet Source	<1%
19	www.arxiv-vanity.com Internet Source	<1%
20	Submitted to IUBH - Internationale Hochschule Bad Honnef-Bonn Student Paper	<1%
21	Ritika Singh, Satwinder Singh. "Text Similarity Measures in News Articles by Vector Space Model Using NLP", Journal of The Institution of Engineers (India): Series B, 2020 Publication	<1%
22	Submitted to University of Witwatersrand Student Paper	<1%
23	cs.christuniversity.in Internet Source	<1%
24	vocal.media Internet Source	<1%
25	Submitted to East West University Center for Research and Training Student Paper	<1%
26	ntnuopen.ntnu.no Internet Source	<1%

27	Goutam Majumder, Partha Pakray, Ranjita Das, David Pinto. "Interpretable semantic textual similarity of sentences using alignment of chunks with classification and regression", Applied Intelligence, 2021 Publication	<1 %
28	David Cornforth, Herbert Jelinek. "Automated classification reveals morphological factors associated with dementia", Applied Soft Computing, 2008 Publication	<1 %
29	K Pushpalatha, V S Ananthanarayana. "An information theoretic similarity measure for unified multimedia document retrieval", 7th International Conference on Information and Automation for Sustainability, 2014 Publication	<1 %
30	Submitted to Cork Institute of Technology Student Paper	<1 %
31	Ramon Ré Moya Cuevas. "A Machine Learning Approach to Portuguese Pronoun Resolution", Lecture Notes in Computer Science, 2008 Publication	<1 %
32	Hou, W.J.. "Enhancing performance of protein and gene name recognizers with filtering and integration strategies", Journal of Biomedical Informatics, 200412 Publication	<1 %

33	arrow.tudublin.ie Internet Source	<1 %
34	etd.gatech.edu Internet Source	<1 %
35	manualzz.com Internet Source	<1 %
36	www.mecs-press.org Internet Source	<1 %
37	ALENA NEVIAROUSKAYA, HELMUT PRENDINGER, MITSURU ISHIZUKA. "Affect Analysis Model: novel rule-based approach to affect sensing from text", Natural Language Engineering, 2010 Publication	<1 %
38	aclanthology.lst.uni-saarland.de Internet Source	<1 %
39	pure.rug.nl Internet Source	<1 %
40	www.iofos.eu Internet Source	<1 %
41	www.sepln.org Internet Source	<1 %