Multilingual News article Similarity

Department of Computer Science and Engineering
National Institute of Technology, Silchar

Presented By (Group No. 49):		
Samudranil Dutta (1912014)	Rajdeep Paul (1912038)	Veeranjaneyulu Chowdary Battula (1912103)
Under the Supervision of : Dr. Partha Pakray		

Outline

1. Introduction
2. Problem Statement
3. Motivation
4. Literature Survey
5. Objectives
6. Proposed Methodology
7. System Prototype
8. Timeline
9. References

Introduction

- Multilingual News Article Similarity is the task of identifying whether news articles address the same subject or not.
- Understanding which articles refer to the same story can not only improve applications like news aggregation but enable cross-linguistic analysis of media consumption and attention.
- This will cover measuring sentence and document similarity.

PROBLEM STATEMENT

To calculate the similarity between two distinct news articles (English-Regional language) and develop a system for determining whether the two news articles are similar or not.

Motivation

Measuring news article similarity enables:

- comparison of news sources coverage
- helps identifying the stories that dominate the media agenda
- tracks the diffusion of news items over time within a media ecosystem.

We found that many people had worked on this previously, but very few covered Indian languages. So we plan to work for English-Indian language pairs and prepare a standard dataset consisting of English-Indian langauge Pairs.

If news articles can be properly clustered, they can be used for a wide range of purposes, such as recommendation and displaying related articles.

Objectives

- To translate the news article from regional language into english language.
- To identify the named entities.
- To calculate the similarity between the two distinct news articles.

Literature Survey

S.No	Authors	Title of the paper	Findings
1	Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, Mattia Samory	SemEval-2022 Task 8: Multilingual news article similarity	Xi et al. measured similarity on basis of following annotators: **GEO, ENT, TIME, NAR, OVERALL, STYLE, TONE**. Errors were particularly common when the news articles shared some similarity in terms of their geographic focus, temporal focus, named entities, and narratives but were nonetheless dissimilar overall.
2	Shotaro Ishihara, Hono Shirai	Nikkei at SemEval-2022 Task 8: Exploring BERT-based Bi-Encoder Approach for Pairwise Multilingual News Article Similarity	The authors used BERT based Bi-Encoder approach for SemEval-2022 task 8. The experiments done in this paper showed that Bi-Encoder architecture worked better than Cross-Encoder.

Literature Survey

S.No	Authors	Title of the paper	Findings
3	Ritika Singh, Satwinder Singh	Text Similarity Measures in News Articles by Vector Space Model Using NLP	They conducted a comparison of three different methods to estimate the semantic similarity among two news articles. The three methodologies are the Cosine similarity with tf-idf vectors, Jaccard similarity with tf-idf vectors, Bag of words Euclidean distance. All three of these methods showed promising results, but among these three methods, cosine similarity using tf-idf showed greater accuracy, recall and F-measure scores of 81.25%, 100% and 76.92%, respectively.

Dataset Preparation

- Prepared a test dataset that contain a pair of news articles.
- Dataset prepared so far consist of 100 pairs.
- The dataset consist of English news and Bengali news as pairs.
- We have manually tagged it as similar or dissimilar.
- 61 pairs were marked as **S** and 39 pairs as **D**.

	Article1 (English)	Article2 (Bengali)	Manually	Bert	roberta
2	To prevent accidents at railway crossings, which constitute as much as 40% of all rail accidents, the Dedicated Freight Corridor will have 1,000 over bridges and underpasses, officials said.	রেল ক্রসিং-এ দুর্ঘটনা প্রতিরোধ করতে উদ্যোগী রেল। আর সেই কারণেই ডেডিকেটেড ফ্রেট করিডোরে(DFC) ১,০০০টি ওভার ব্রিজ এবং আন্ডারপাস থাকবে বলে জানালেন রেলের আধিকারিকরা।	S	0.957999	0.888648
3	An early Sunday shooting outside a bar in downtown Tampa, Florida has left one person dead and six wounded, police said. The Tampa Police Department said in a news release that the	পরিবারের লোকেদের নামে একাধিক ব্যাঙ্ক অ্যাকাউন্টের পাশাপাশি অজ্ঞাতপরিচয় ব্যক্তিদের নামেও অ্যাকাউন্ট খুলেছিলেন মানিক ভট্টাচার্য। পলাশিপাড়ার তৃণমূল বিধায়কের নামে আরও গুরুতর অভিযোগ এনফোর্সমেন্ট ডিরেক্টরেটের	D	0.55807	0.105500

Methodology

We shall be doing three things:

Translation:

- The regional language is translated to English language.
- It is achieved by using the microsoft bing translate api which converts the given text of a language into target language.

Named Entity Recognition:

- We find out the named entities from the given text and classify them into different categories.
- Example of categories include Person, Organization, Place/Location, etc.
- Stanza and corenlp API are used for doing the same.
- Core NLP is able to recognize named entities better than stanza.

Calculating Similarity:

- To find out the similarities between the two articles we have so far used cosine function and BERT+Cosine similarity.
- In case of Cosine similarity, the text is first converted to vectors and the similarity of two vectors are 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.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Methodology (contd…)

BERT:

- Bidirectional Encoder Representations from Transformers.
- BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task.
- As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. Therefore it is considered bidirectional, though it would be more accurate to say that it's non-directional. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word).

Methodology (contd…)

RoBERTa:

- Robustly optimized BERT approach RoBERTa, is a retraining of BERT with improved training methodology, 1000% more data and compute power.

- To improve the training procedure, RoBERTa removes the Next Sentence Prediction (NSP) task from BERT's pre-training and introduces dynamic masking so that the masked token changes during the training epochs.

- Additional data included from CommonCrawl News dataset (63 million articles, 76 GB), Web text corpus (38 GB) and Stories from Common Crawl (31 GB)

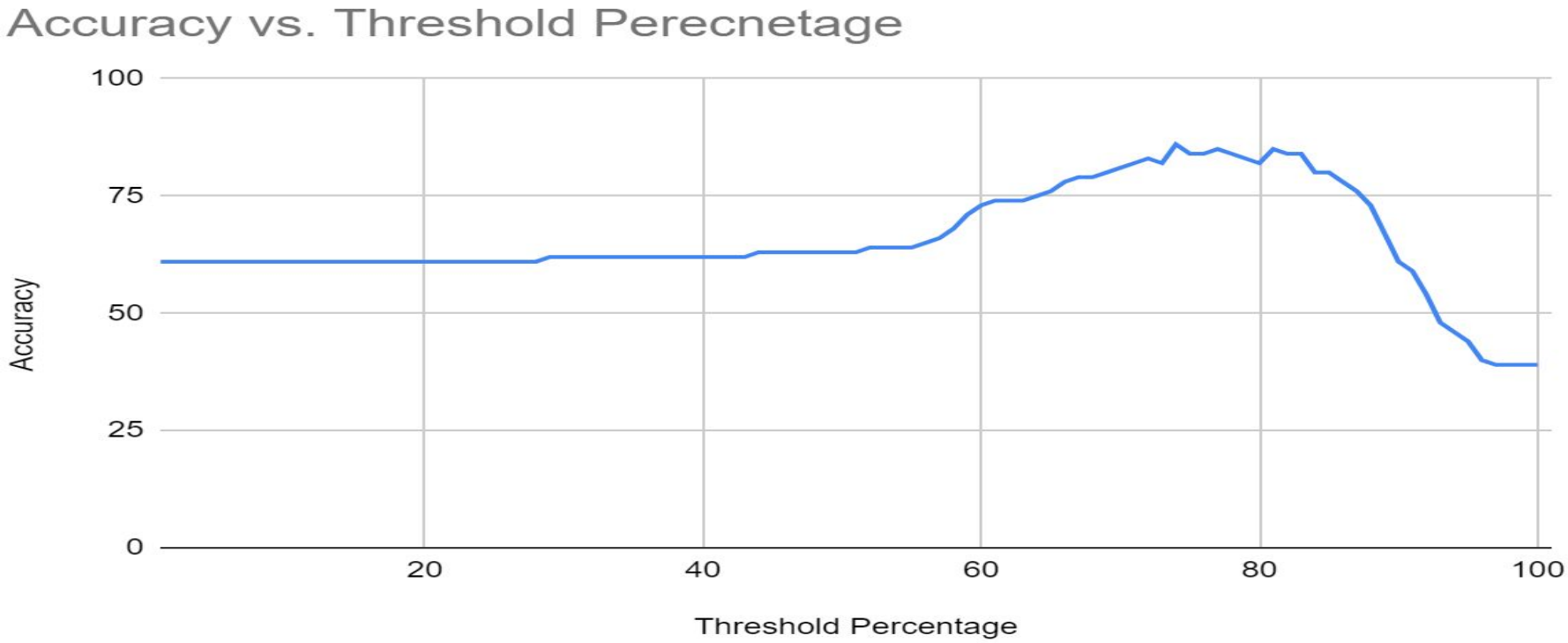
Evaluation Results

- We tested the bert and roberta model on the sample 100 test cases. The result (percentage) we obtained were used to calculate the threshold value.
- If the result obtained is below the threshold value then the articles will be considered as dissimilar. Else it will be considered as similar.
- To calculate the threshold value we plotted accuracy vs threshold value graph.

Accuracy= Number of correct predictions/Total number of predictions

Evaluation Results (Bert)

Threshold value calculation



Evaluation Results (Bert)

- From the graph we found that accuracy is maximum when the threshold is kept at 74%. Accuracy is 86.
- Keeping this threshold value as reference, we evaluated the result obtained from testing our test data on our model.
- We calculated the precision, recall and the F-score.

Precision, Recall and F-score (Bert)

	ACTUAL SIMILAR	ACTUAL DISSIMILAR
SYSTEM TAGGED SIMILAR	60(TP)	13 (FP)
SYSTEM TAGGED DISSIMILAR	1(FN)	26(TN)

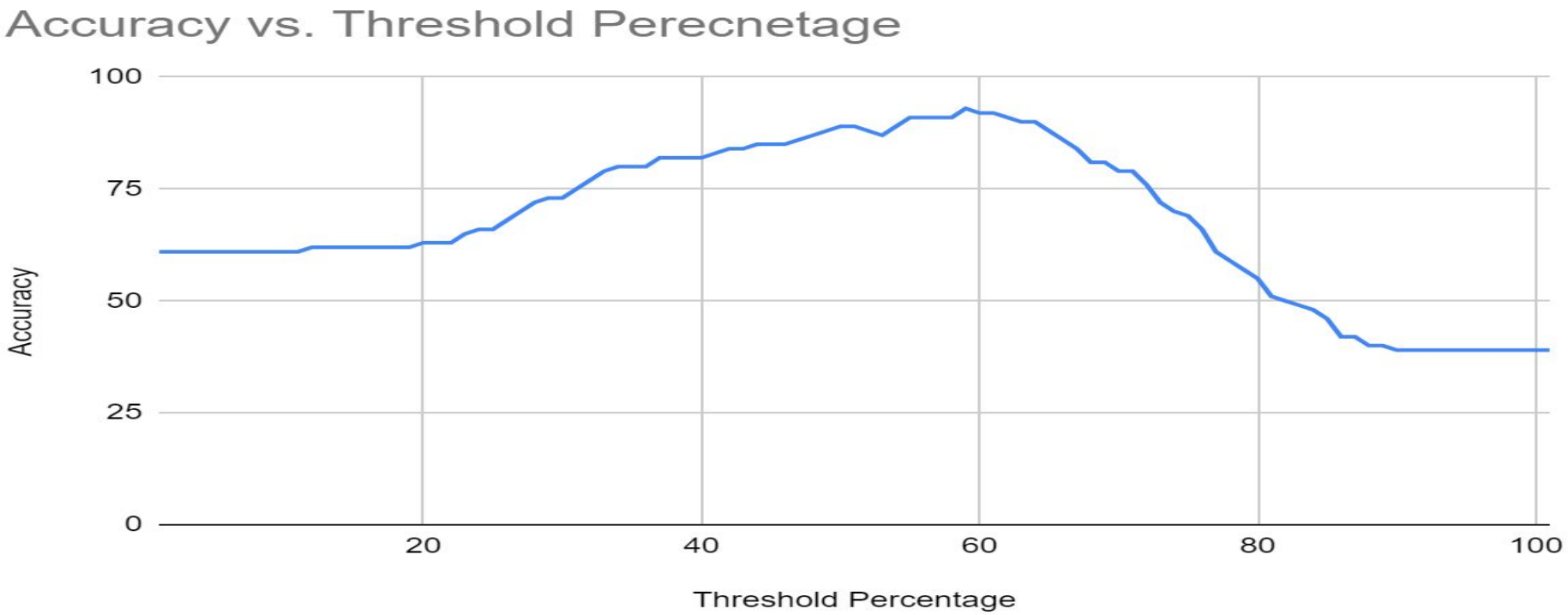
PRECISION= TP/(TP+FP)= 60/(60+13)= 0.821

RECALL =TP/(TP+FN)= 60/(60+1)= 0.983

F-SCORE =2PR/(P+R) = 0.894

Evaluation Results (Roberta)

Threshold value calculation



Evaluation Results (Roberta)

- From the graph we found that accuracy is maximum when the threshold is kept at 59%. Accuracy is 92.
- Keeping this threshold value as reference, we evaluated the result obtained from testing our test data on our model.
- We calculated the precision, recall and the F-score.

Precision, Recall and F-score (RoBERTa)

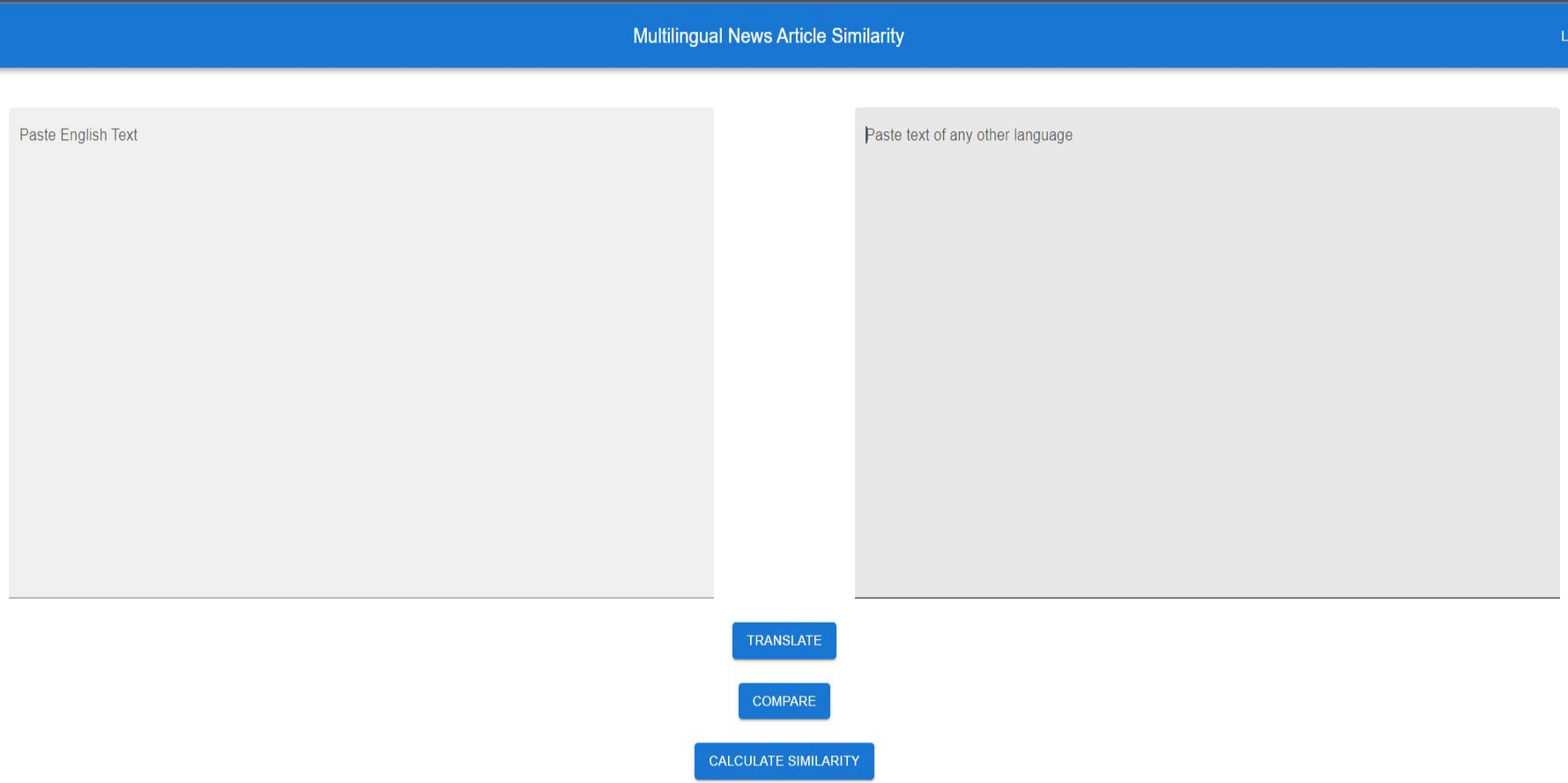
	ACTUAL SIMILAR	ACTUAL DISSIMILAR
SYSTEM TAGGED SIMILAR	56 (TP)	3 (FP)
SYSTEM TAGGED DISSIMILAR	5 (FN)	36 (TN)

PRECISION= TP/(TP+FP)= 56/(56+3)= 0.949

RECALL =TP/(TP+FN)= 56/(56+5)= 0.918

F-SCORE =2PR/(P+R) = 0.933

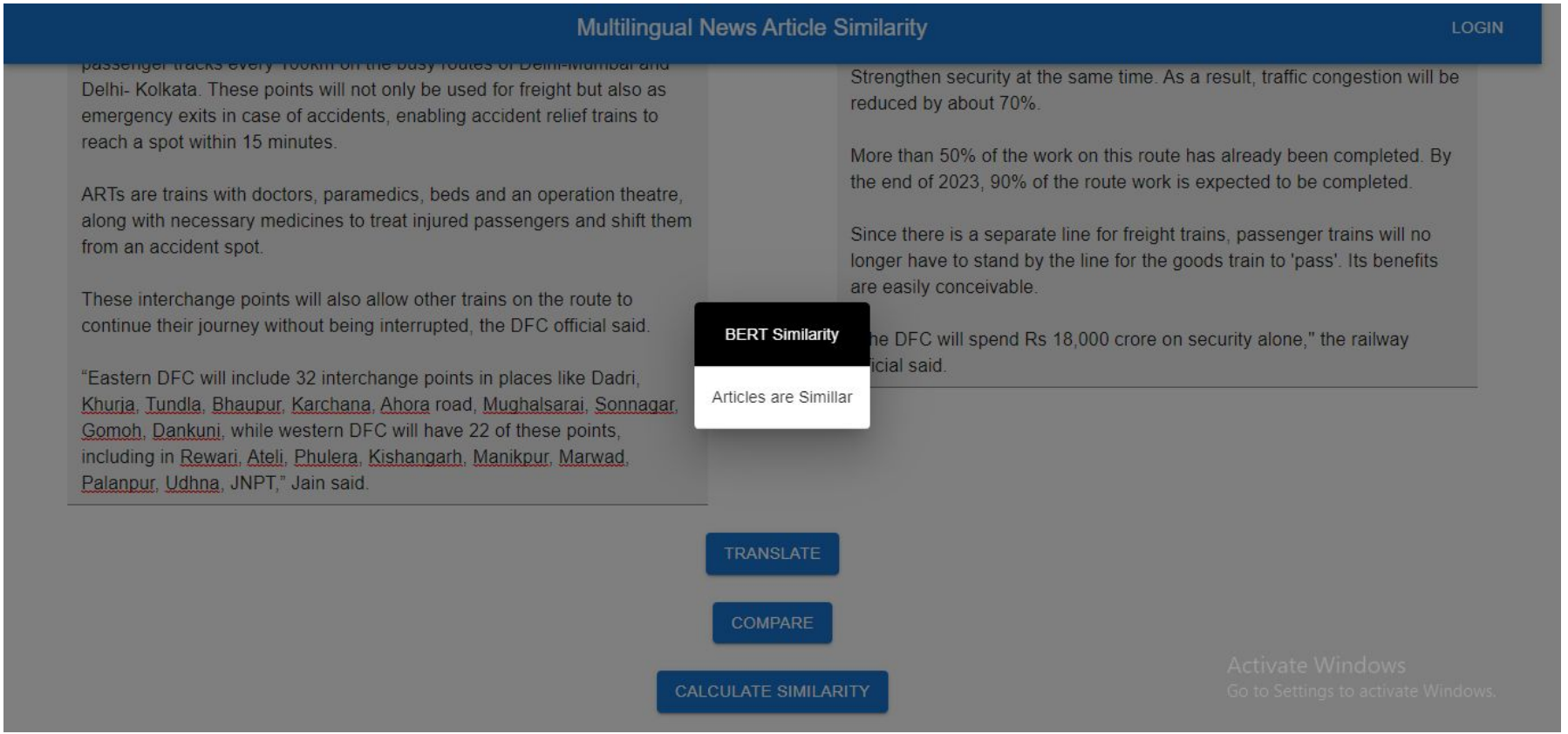
System Prototype



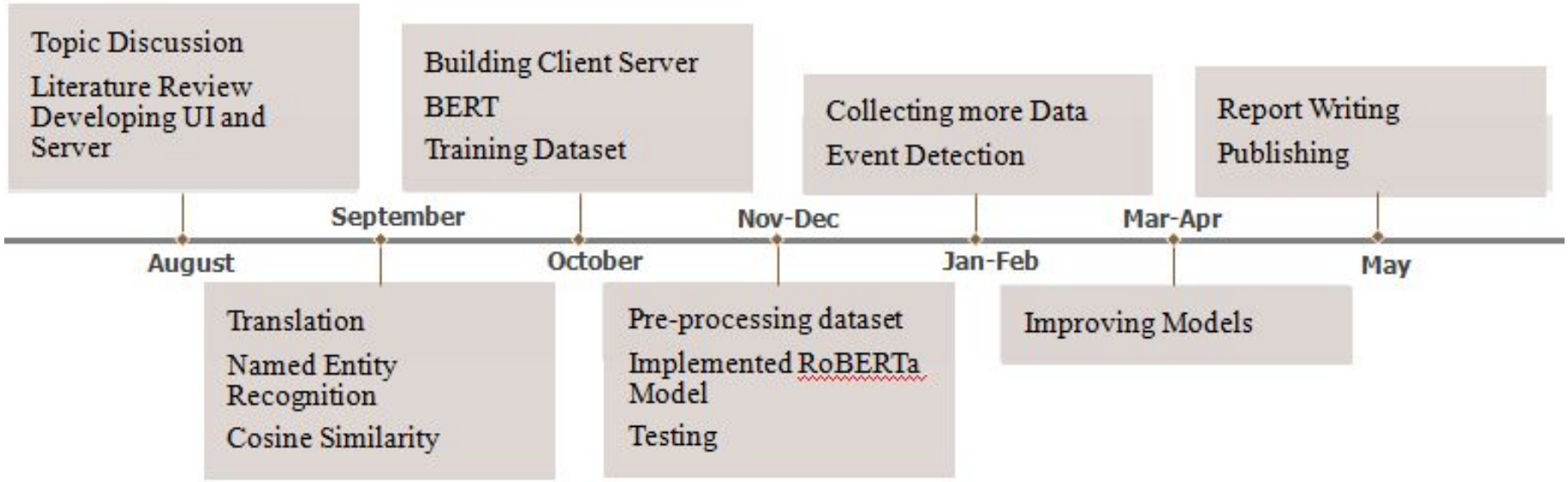
System Glimpse

Parameters	English Language	Regional Language
CITY	• Amravati	• Nabha • Mumbai
DATE	• 2022-09-11 • PAST_REF	• 2022-09-08
DURATION	• 35 years old	• the last few days • the last few days
LOCATION	• Maharashtra	• Seva • Navi
NATIONALITY	• Indian	
ORGANIZATION	• BJP • Amravati Police	• the Directorate of Revenue Intelligence • DRI • DRI • DRI • DRI • DRI
PERSON	• Nupur Sharma	• Nabha Sewa • Nabha Sewa
TITLE	• chemist • Prophet	
COUNTRY		• South Africa • South Africa

System Glimpse



Timeline


Topic Discussion
Literature Review
Developing UI and Server
Building Client Server
BERT
Training Dataset
Collecting more Data
Event Detection
Report Writing
Publishing
August
September
October
Nov-Dec
Jan-Feb
Mar-Apr
May
Translation
Named Entity Recognition
Cosine Similarity
Pre-processing dataset
Implemented RoBERTa Model
Testing
Improving Models

References

[1] Nikkei at SemEval-2022 Task 8: Exploring BERT-based Bi-Encoder Approach for Pairwise Multilingual News Article Similarity

[2] SemEval-2022 Task 8: Multilingual news article similarity

[3] The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents D Gunawan1, C A Sembiring and M A Budiman

[4] Overview - CoreNLP (stanfordnlp.github.io)

[5] Microsoft Bing API translator

[6] BERT Explained: A Complete Guide with Theory


Thank you!