# Multilingual News Article Similarity

Arnab Ghorui, Jayadratha Gayen, Puru Ojha, Rudra Dhar

Team: Nearest Neighbours (54)

Major project for Statistical Methods in AI

**Abstract**

This report describes our project for the course SMAI. Here we made a system to determine multilingual news article similarity. Specifically, we tackled task 8 of SemEval-2022. The task of multilingual news article similarity entails determining the degree of similarity of a given pair of news articles in a language-agnostic setting. We used a Siamese Architecture, as it's a well-known method for text comparison tasks. We developed 3 models for the task. In our baseline model we used tf-idf for encoding for MLP's for the model architecture. We developed another model where we used Glove vectors for encoding, and LSTM and MLP for modeling. Lastly we made a transformer based architecture, where we used multilingual DistilBERT. We got the best results with Bert with a score of 0.2973.

**Keywords:** news similarity; tf-idf; MLP; Glove; LSTM; transformers; DistilBERT;

# 1 Introduction

Numerous media outlets publish thousands of new news articles every day. Understanding which articles refer to the same topic not only improves applications such as news aggregation but also allows for cross-linguistic research of media consumption and attention. Due to the various ways in which a story might differ, such as when two pieces have a lot of textual overlap but depict the same events that occurred years apart, or when there is very little textual overlap but the news stories talk about the same event, determining how similar two news articles are can be difficult. Therefore we have chosen to determine the similarity between news articles in a multilingual setting. Particularly we have attempted *SemEval 2022 Task 8: Multilingual News Article Similarity* (2022) organized by Chen et al. (2022) . Here they have more than 7000 news article pairs in 18 languages, which we had to score based on their similarity.

We have used Siamese Architecture to predict the similarity score, as it's a well-known method for text comparison tasks. We used 3 main architectures; the first uses tf-idf for encoding and MLP for regression score; the 2nd uses Glove for word encoding, LSTM and MLP for regression score; and in the final model we used transformer encoders for encoding, and dense layer for regression score.

We referred the works of Joshi, Taunk, and Varma (2022) for Siamese Architecture and Xu, Yang, Cui, and Chen (2022) for multilable learning, and came up with several improved models.

We have made our code available on GitHub.[1]

# 2 Task and Dataset description

## 2.1 Task

This is a Sem-Eval 2022 task. In this task the input will be a pair of newspaper articles in the same or different languages and, the similarity had to be determined as to whether the news articles are of the same topic i.e. to calculate the similarity scores between the news articles on a 4-point scale from least to most similar.

## 2.2 Dataset

For the training dataset, we are given the language of the news articles, a pair id for the news articles, the corresponding links of the news articles to extract the data, & 7 types of similarity scores based on the different information that can be extracted from the data(ex: tone of the articles, location where the incident took place, the time when the incident happened). Then there is the overall similarity score which is based on all the factors in the article. The training dataset contains the language pairs ar-ar, ar-en, de-de, de-en, en-en, en-es, en-fr, en-pl, es-es, fr-fr, pl-pl, tr-tr and the size of the dataset is 4918 out of which only 2857 were with the valid links

[1] https://github.com/arnabghorui/smai_project

from which the data could be extracted. The test dataset contained the language pairs ar-ar, de-de, de-en, en-en, en-es, en-pl, es-es, fr-fr, pl-pl, and tr-tr, and in addition, the language pairs de-fr, de-pl, es-it, fr-pl, it-it, ru-ru, zh-en, zh-zh which are not present in the training dataset. The size of the test dataset is 4902 out of which 4316 were with valid links from which the data could be extracted. With the links given in the data, the information was extracted, and then it was subsequently cleaned so that it can be used for training and testing purposes.

# 3 System Description

We used a Siamese based architecture to detect the similarity between two news articles. For MLP and LSTM we have used tensorflow to make the model whereas in DistilBERT we have used Pytorch.
The basic Siamese architecture used for the different models is as shown below. For MLP the embedding is tf-idf, for LSTM it is GloVe, for transformers it is DistilBERT encoding.

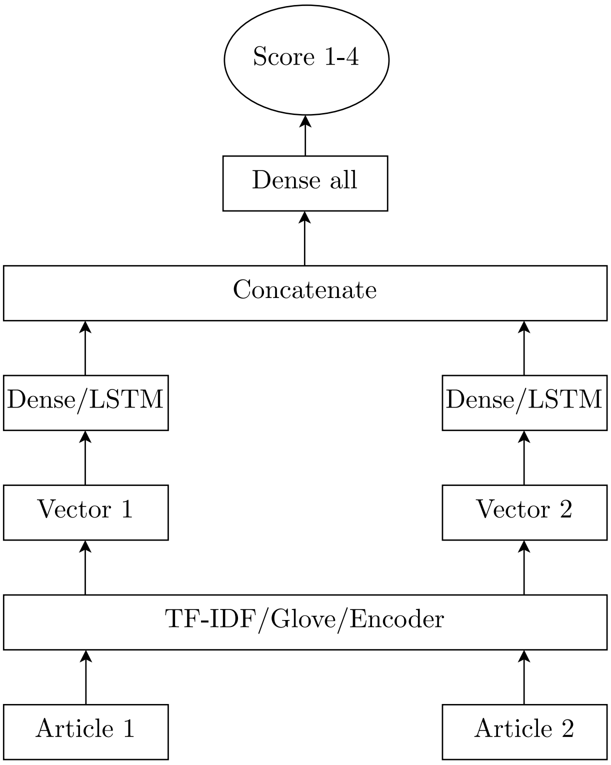

Figure 1: Siamese architecture

We made various models using MLP, LSTM, and Transformer Encoder as described below:

## 3.1 Multilayer Perceptron using tf-idf:

TF-IDF is a statistical measure used to determine the mathematical significance of words in documents. The TF-IDF value is obtained by multiplying Total Frequency (TF) of a word in a given article to Inverse Document Frequency (IDF) from the corpus. Multilayer Perceptrons are a class of fully connected, feed forward Artificial Neural Networks. We use TF-IDF embedding to perform regression via a Multilayer Perceptron Network. Once fully learned this model is then used to check the similarity between two articles.

## 3.2 LSTM using glove encodings:

Global Vector (GloVe) [Pennington, Socher, and Manning (2014)] is a pretrained embedding available on the web. It is an unsupervised learning algorithm developed by Stanford for generating word embeddings by aggregating global word-word co-occurrence matrix from a corpus. Long Short Term Memory networks (LSTM) are a special type of Recurrent Neural Networks which are capable of learning long term dependencies and Bidirectional LSTM are those LSTM's in which the information can flow in both directions backwards and forward. We have used GloVe encoding to train LSTM Networks to find the similarity score.

### 3.3 Transformer - DistilBERT

Transformers are the current State of The Art in the field of Natural Language Processing. They are designed to take sequential data as input and are able to focus on a particular part of the input that is relevant to make the current prediction. BERT[Devlin, Chang, Lee, and Toutanova (2018)] is a transformers model pretrained on a large corpus of English data in a self-supervised fashion. We used 'distilbert-base-cased' as Encoder for producing the similarity score, for the English article data. For the data with articles in all languages we used 'distilbert-base-multilingual-cased'.

### 3.4 Multilable Learning

The Training data had extra labels for score like geography and time which we didn't need to predict. But these labels can be used for training. So we designed a experiment where we trained with these 7 labels, but tested only with the OVERALL score as given in the task.

## 4 Experiments and Results

### 4.1 Evaluation Criteria

Since the task is regression, following metrics are used

- **PCC** The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. It is defined as follows.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- **MSE** Mean Squared Error:

$$MSE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2$$

### 4.2 Experiments Performed

1. **MLP_TI :** First we filter out all the pairs containing English articles. We then used TF-IDF for encoding the news articles. We concatenated the two encodings as [X1, X2] and then used a Dense layer for the Modeling the regression. Here only overall score is used for the training and testing.
2. **MLP_TI_1 :** After filtering out the English articles We used TF-IDF for encoding the news articles. Here we concatenated the two encodings as [X1+X2, X1-X2] and then used a single Dense layer for the Modeling. Overall score is used for the training and testing.
3. **MLP_TI_MultiLable :** We used the MLP_TI_1 model as described above. In place of the overall score we used multi-label loss as described in 3.4
4. **MLP_TI_MultiLingual :** We used the MLP_TI_1 model as described above. But this experiment was a done in MultiLingual setting on all data (not just English)
5. **LSTM_G :** After filtering out the English articles We used glove encoding for encoding the words. Next LSTM's were used to for encoded articles. We concatenated the two encodings as [X1, X2]. Finally a dense layer was used to get the similarity score.
6. **LSTM_G_1 :** In continuation with the model LSTM_G in place of normal LSTM Bi-LSTM's were used to for encoding the sentences. Here We concatenated the two encodings as [X1+X2, X1-X2]. Finally a dense layer was used to get the similarity score.
7. **LSTM_G_MultiLable :** We used the LSTM_G_1 model as described above. Along with that in place of Overall score we used multi-label loss for training as described in 3.4
8. **DB :** For the English Articles from the dataset We used DistilBERT-base-cased as Encoder for encoding the news articles. We concatenated the two Encoder outputs as [X1+X2, X1-X2]. Finally a dense layer was used to get the similarity score.
9. **DB_M_Lingual :** We used the DB model as described above. But this experiment was a done in MultiLingual setting on all data (not just English)
10. **DB_M_Lable :** We used the DB_M_Lingual model as described above. Along with that in place of just Overall Score we used multi-label loss for training as described in 3.4

### 4.3 Results:

Table 1: Results of all the experiments

| EXPERIMENT | TRAIN MSE | DEV MSE | TEST MSE | TEST PCC |
|---|---|---|---|---|
| MLP_TI | 0.0183 | 1.7104 | 1.8702 | 0.1909 |
| MLP_TI_1 | 0.0259 | 1.1287 | 1.8045 | 0.2465 |
| MLP_TI_M_Label | 0.0250 | 1.5680 | 0.5113 | 0.5683 |
| MLP_TI_M_Lingual | 0.1015 | 1.3703 | 1.6572 | 0.0148 |
| LSTM_G | 1.0378 | 1.6435 | 1.9589 | 0.1624 |
| LSTM_G_1 | 0.7773 | 1.9335 | 1.9891 | 0.2591 |
| LSTM_G_M_Label | 0.2610 | 1.7065 | 1.3553 | 0.4626 |
| DB | 0.0212 | 0.9833 | 1.4693 | 0.3561 |
| DB_M_Lingual | 0.0854 | 1.203 | 1.3172 | 0.2817 |
| DB_M_Label | 0.0922 | 1.146 | 1.4209 | 0.2389 |

## 5 Observations and Conclusion

The results show us that News similarity prediction is much difficult in Multilingual setting than in Uni-lingual (English) setting. This is clear from the MSE and PCC scores in the test data. Even the train loss is much greater in the Multilingual setting.

We can also observe that the Multi-label Learning gives better performance on average in test data, as it generalizes better. We make this observation on MLP and LSTM based models, but crucially not on BERT based model. Though it is important to note that the multi-label experiment done on BERT is also on multilingual data.

We observe from the MLP and the LSTM based experiments (MLP_TI_1 and LSTM_G_1) that the model performs better when we do the concatenation as [X1+X2, X1-X2] rather than [X1, X2] (where X1 and X2 are the encodings for each of the news articles in pairs). So [X1+X2, X1-X2] concatenation was used for the multilable and multi-lingual experiments. In our final experiment using Bert we only used the [X1+X2, X1-X2] concatenation (where X1 and X2 are encoding output of the DistilBERT Encoder)

PCC was given as one of the key metrics. But as it's a regression problem we used a MSE as loss functions in all our models. But it is crucial to note the PCC doesn't vary (inversely) with MSE always as is evident from the results.

Overall, we attempted the problem of Multilingual News Similarity Detection, organised by SemEval 2022 task8. Our basic model architecture was Siamese based. We used three different acrtitecture, MLP based, LSTM based, and transformer based, and performed many different experiments. Overall heavier and more complex model gave better results. We got the best result on unilingual setting with MLP using Multilable training where the test MSE was 1.31 , and PCC was 0.2817. We got the best result on multilingual setting with DistilBERT where the test MSE was 0.5113 , and PCC was 0.5683.

## References

Chen, X., Zeynali, A., Camargo, C., Flöck, F., Gaffney, D., Grabowicz, P., . . . Samory, M. (2022, July). SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th international workshop on semantic evaluation (semeval-2022)* (pp. 1094–1106). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.semeval-1.155 DOI: 10.18653/v1/2022.semeval-1.155

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding.* Retrieved from http://arxiv.org/abs/1810.04805 (cite arxiv:1810.04805Comment: 13 pages)

Joshi, S., Taunk, D., & Varma, V. (2022, July). IIIT-MLNS at SemEval-2022 task 8: Siamese architecture for modeling multilingual news similarity. In *Proceedings of the 16th international workshop on semantic evaluation (semeval-2022)* (pp. 1145–1150). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.semeval-1.161 DOI: 10.18653/v1/2022.semeval-1.161

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*

*(EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D14-1162 DOI: 10.3115/v1/D14-1162

*Semeval 2022 task 8: Multilingual news article similarity.* (2022). Retrieved from https://competitions.codalab.org/competitions/33835

Xu, Z., Yang, Z., Cui, Y., & Chen, Z. (2022, July). HFL at SemEval-2022 task 8: A linguistics-inspired regression model with data augmentation for multilingual news similarity. In *Proceedings of the 16th international workshop on semantic evaluation (semeval-2022)* (pp. 1114–1120). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.semeval-1.157 DOI: 10.18653/v1/2022.semeval-1.157