Report of Task 3: Assessing multilingual news article similarity

Name: Aditya Brahme

1. Introduction

Digital media has grown exponentially in recent years, and as a result, enormous amounts of data are produced daily. A variety of content, including news articles in several languages, is included in this data. It is essential to create tools that can quickly process and analyze this large amount of data in order to understand and make sense of it. News articles from different sources and in different languages may cover similar or related topics, and it is important to identify these similarities in order to gain a better understanding of the overall picture.

The effort of comparing and evaluating the similarity of news stories across several languages is known as the problem of assessing multilingual news article similarity. It is crucial to create effective methods for comparing and evaluating the similarity of news stories across many languages as a result of the growing amount of digital material produced in various languages.

The goal of this problem is to identify news articles that are semantically similar, regardless of the language in which they were written. The semantical similarity can be used to enhance the performance of the search engines. Document similarity can be used to detect instances of plagiarism, where one document is copied from another without proper attribution. Additionally, it can be used to recommend related or similar documents to users.

Challenges in this problem are variability in language (Documents in different languages can vary greatly in terms of grammar, syntax, and vocabulary), ambiguous and subjective (interpreted differently by different people), scale (amount of text data grows) and context of the articles, domain specific vocabulary.

2. Problem Formulation

Assessing multilingual news article similarity is the task of determining the degree of similarity between two news articles written in different languages. The task can be formally defined as follows:

Input: 2 news articles written in different languages.

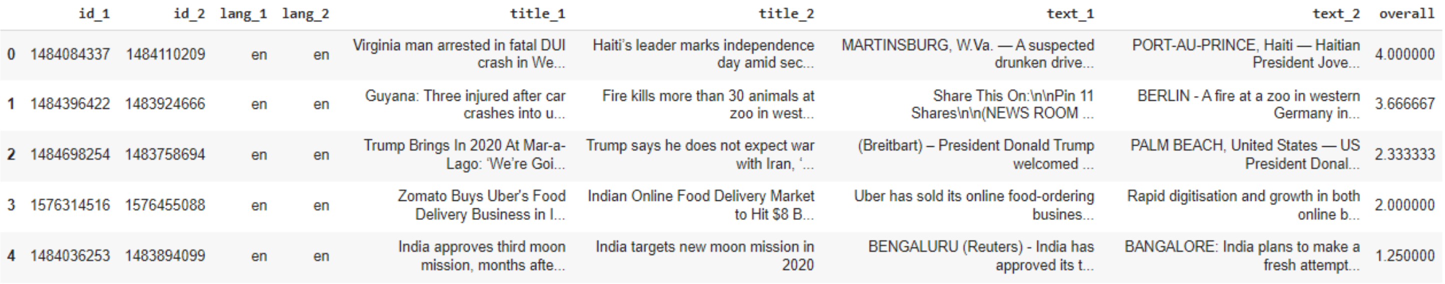
Output: A similarity score between the input articles that has float values between 1 – 4. 1 indicates least similar and 4 indicates closely similar.

Task Type: It is a regression problem. Where the model will need to predict the output values ranging between 1 – 4. It is not a classification problem as there are no specific classes or groups.

3. Method

3.1 Load Data:

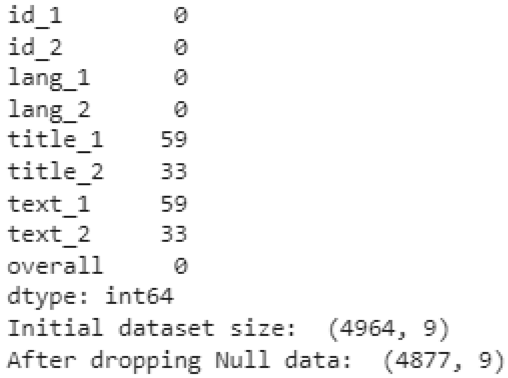
- Step 1: Load json files from Google Drive
- Step 2: Load csv file for train and test dataset
- Step 3: Map the ids from the csv file to json file name and store the pairs in a pandas dataframe. Following is the structure of the dataframe create



	id_1	id_2	lang_1	lang_2	title_1	title_2	text_1	text_2	overall
0	1484084337	1484110209	en	en	Virginia man arrested in fatal DUI crash in We...	Haiti's leader marks independence day amid sec...	MARTINSBURG, W.Va. — A suspected drunken drive...	PORT-AU-PRINCE, Haiti — Haitian President Jove...	4.000000
1	1484396422	1483924666	en	en	Guyana: Three injured after car crashes into u...	Fire kills more than 30 animals at zoo in west...	Share This On:\n\nPin 11 Shares\n\n(NEWS ROOM ...	BERLIN - A fire at a zoo in western Germany in...	3.666667
2	1484698254	1483758694	en	en	Trump Brings In 2020 At Mar-a-Lago: 'We're Goi...	Trump says he does not expect war with Iran, '...	(Breitbart) – President Donald Trump welcomed ...	PALM BEACH, United States — US President Donal...	2.333333
3	1576314516	1576455088	en	en	Zomato Buys Uber's Food Delivery Business in I...	Indian Online Food Delivery Market to Hit $8 B...	Uber has sold its online food-ordering busines...	Rapid digitisation and growth in both online b...	2.000000
4	1484036253	1483894099	en	en	India approves third moon mission, months afte...	India targets new moon mission in 2020	BENGALURU (Reuters) - India has approved its t...	BANGALORE: India plans to make a fresh attempt...	1.250000

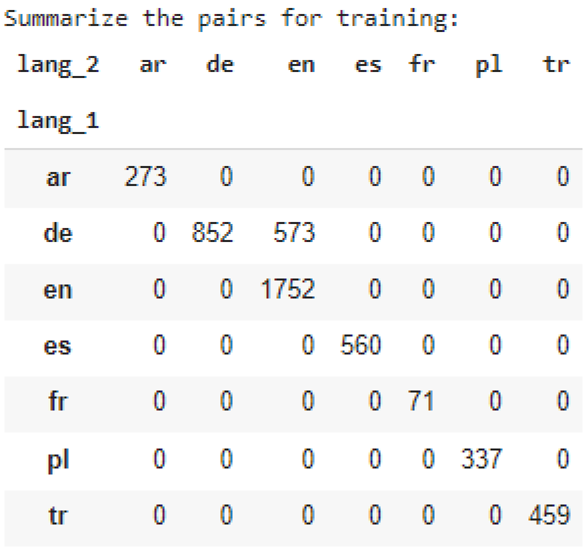
3.2 Data Pre-processing:

- Check for Null entries: Rows having null entries are being removed. The following table shows the number of null values in train dataset



```
id_1        0
id_2        0
lang_1      0
lang_2      0
title_1    59
title_2    33
text_1     59
text_2     33
overall     0
dtype: int64
Initial dataset size:  (4964, 9)
After dropping Null data:  (4877, 9)
```

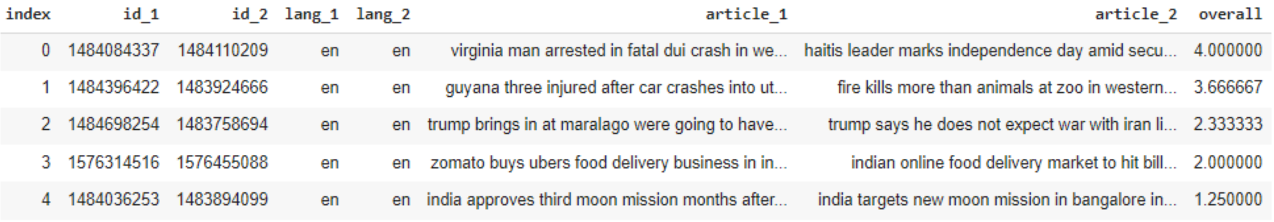
- Clean data: In this step we will remove URLs, special characters, tabs, break lines, digits and convert the text to lowercase format.
- Normalize the output: As the output score varies between 1-4, MinMaxScaler is used to transform the output vector in range of 0-1. So that in the model we can use sigmoid as the output activation function.



Summarize the pairs for training:

lang_2	ar	de	en	es	fr	pl	tr
lang_1							
ar	273	0	0	0	0	0	0
de	0	852	573	0	0	0	0
en	0	0	1752	0	0	0	0
es	0	0	0	560	0	0	0
fr	0	0	0	0	71	0	0
pl	0	0	0	0	0	337	0
tr	0	0	0	0	0	0	459

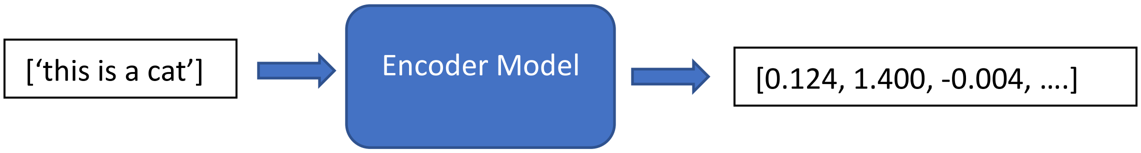
Final dataframe:



index	id_1	id_2	lang_1	lang_2	article_1	article_2	overall
0	1484084337	1484110209	en	en	virginia man arrested in fatal dui crash in we...	haitis leader marks independence day amid secu...	4.000000
1	1484396422	1483924666	en	en	guyana three injured after car crashes into ut...	fire kills more than animals at zoo in western...	3.666667
2	1484698254	1483758694	en	en	trump brings in at maralago were going to have...	trump says he does not expect war with iran li...	2.333333
3	1576314516	1576455088	en	en	zomato buys ubers food delivery business in in...	indian online food delivery market to hit bill...	2.000000
4	1484036253	1483894099	en	en	india approves third moon mission months after...	india targets new moon mission in bangalore in...	1.250000

3.3 Sentence embedding:

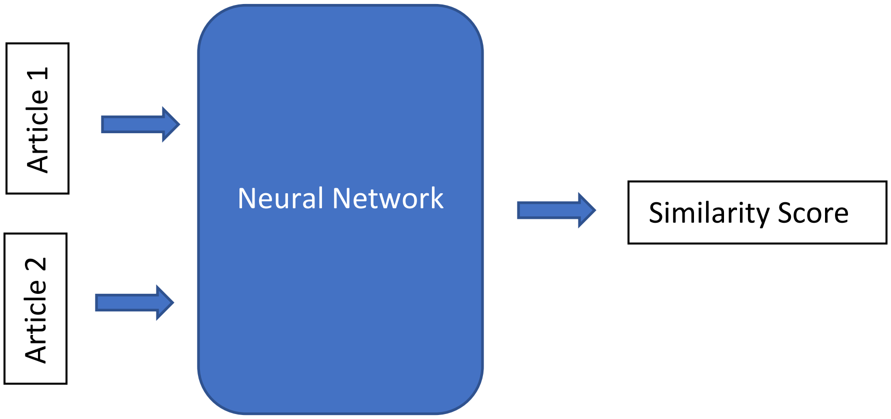
- SentenceTransformers is a Python framework for state-of-the-art sentence, text and image embeddings. The initial work is described in paper Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.
- The Hugging Face team created the pre-trained transformer-based model sentence-transformers/paraphrase-multilingual-mpnet-base-v2 to produce high-quality sentence embeddings. The model is built on a large-scale transformer model called the Megatron-Pretrained Neural Network (MPNet), which was trained on a substantial amount of multilingual text data.
- The paraphrase-multilingual-mpnet-base-v2 model is particularly trained to produce semantically relevant embeddings in more than 50 languages
- The model architecture is based on the transformer encoder-decoder architecture, which incorporates multiple layers of self-attention mechanisms and feedforward neural networks. The model has 12 transformer layers with a hidden size of 768 and a total of 162M parameters.



Dimension of output vector is: (1, 768)

3.4 Models:

- The embedding model has encoded all the articles from different languages in a common vector space. Next, we will give these sentences embedding as an input to a simple neural network so that it can learn to find similar between two articles.



Each article is of dimension of (1,768)

Loss Function: Mean Absolute Error (MAE)

$$\textbf{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Output Activation Function: Sigmoid Function

3.5 Evaluation:

On the final model the MAE score for test dataset is **0.30576**

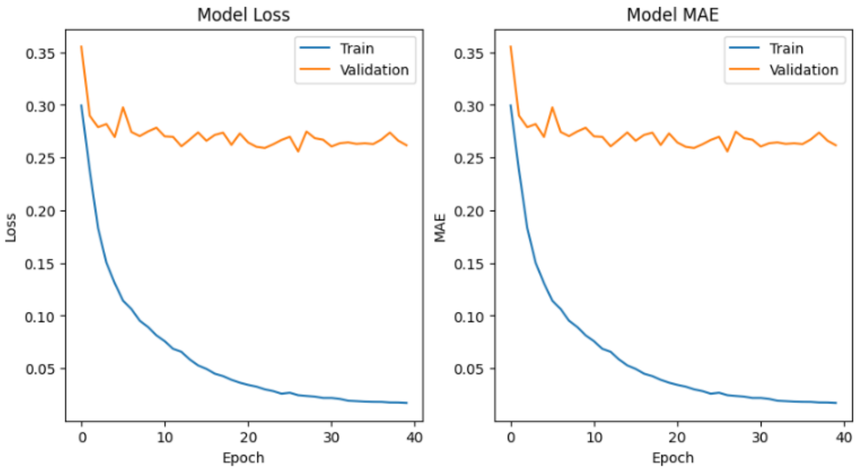
4. Experiments

Implemented 4 models with different hidden layers, learning rate, epochs and regularization techniques. All the model have MAE as the metric as well as the loss function.

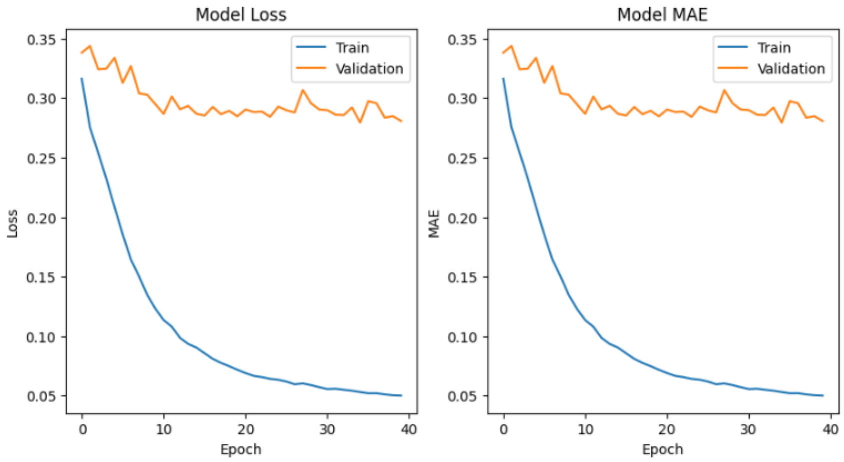
	Model 1	Model 2	Model 1 – regularization	Model 2 - regularization
Number of hidden layers	4	3	4	3
Optimizer & Learning Rate	Adam (lr = 0.0005)	Adam (lr = 0.0005)	Adam (lr = 0.0005)	Adam (lr = 0.0001)
Epochs	40	40	60	100
Regularization	None	None	L2	L2
Dropout	None	None	Yes	Yes
Batch Normalization	None	None	Yes	Yes

Loss and MAE Graphs:

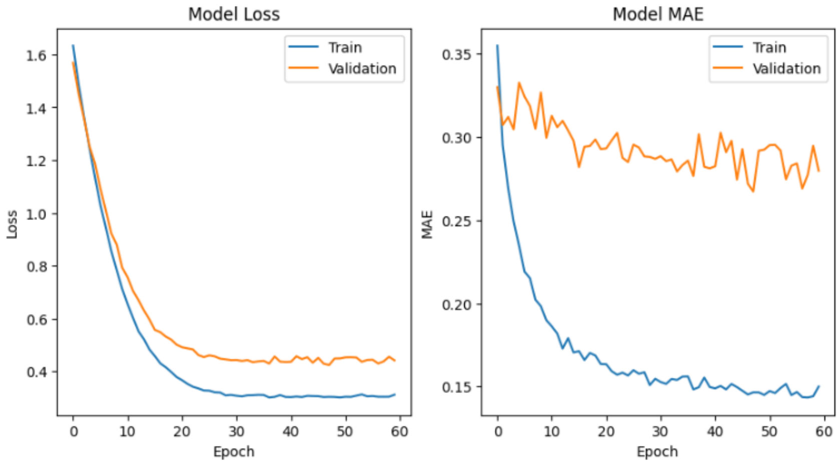
Model 1:



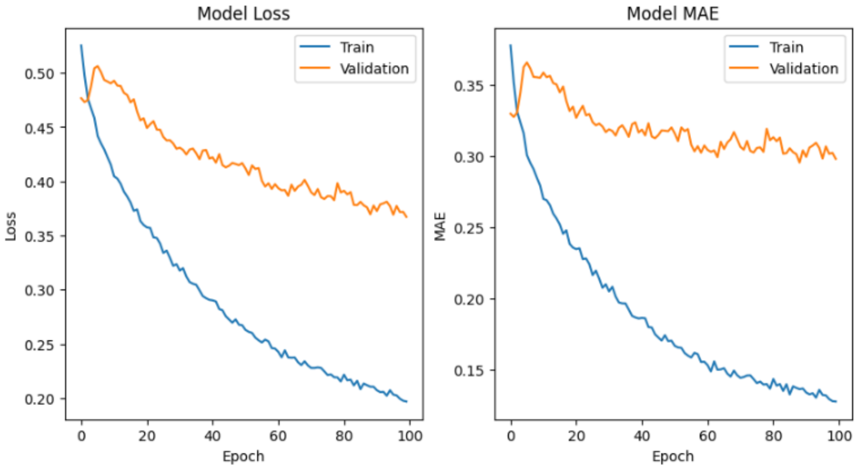
Model 2:



Model 1 – regularization:



Model 2 – regularization:



Model 1 and Model 2 are overfitting.

In Model 2 – regularization, loss is decreasing as well as the MAE score is also decreasing. We can say this model is learning better compared to other models.

5. Conclusion

Assessing multilingual news article similarity is a challenging problem in NLP because you will need to deal with variety of languages. SentenceTransformers embedding model has a significant role in converting all the articles into a common vector space. Using which I was able to build a fully connected neural network to find the similarity score.

The Final MAE score on Testing dataset is 0.30576

Future score: To improve the performance of the model we can augment the training data and use it for training the model. As you can observe we do not some of the language pairs in our training dataset. For augmenting the data, we can use Google Translator API.