Detecting anomalies, contradictions, and contextual analysis through NLP in text

Shivam Sharma
School of Computer Science Engineering and Technology
Bennett University
Greater Noida, India
sharma.shivamhr@gmail.com

Mirdul Swarup
School of Computer Science Engineering and Technology
Bennett University
Greater Noida, India
mirdul2610@gmail.com

Tanush Mahajan
School of Computer Science Engineering and Technology
Bennett University
Greater Noida, India
tanush.mohina@gmail.com

Zeel Dilipkumar Patel
School of Computer Science Engineering and Technology
Bennett University
Greater Noida, India
patelzeel68@gmail.com

2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT) | 978-1-6654-8268-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICT55121.2022.10064560

***Abstract*— This work aims to find contradictions between various sentence pairs, practical application of the model involves various disciplines such as law enforcements for comparison of various witness testimonies or to understand historical manuscripts from different point of views, using NLP based model training such as neural networks, KNN and random forest. We would try to find out the model providing us with the most accurate figures suitable.**

Keywords—Natural Language Processing, Random Forest, RoBERTa, AdaBoost, KNN, tokenizer.

I. INTRODUCTION

A text can easily contain a lot of discrepancies, inconsistencies and errors in itself which makes the use of a context analyzer and contradiction detector more useful. A collection of contradicting facts that mislead consumers is a common component of disinformation and when the amount of information is so enormous that it is unmanageable, the ability to recognize contradictory information immediately becomes critical [9]. Suppose there are two simple statements which are said by two different individuals on an argument: "I am in favour of this resolution" and the other says "I am not in favour of this resolution", this servers as a simple example of two statements which could be run through a context analyzer to determine if there's a contraction in it or not, while these statements are fairly simple, there could be statements with use of more antonyms and synonyms which might require a deeper understanding of the language to determine a certain context through contradiction.

The intense nature of this topic itself opens up doors to a lot of real-world applications in general, a lot of real-world applications have been worked on this topic and few of which: Inconsistency recognition has become one alongside the content procedures, Sanda M. and Andrew encouraged with their project [1], like series of questions and multiple documents concise overview mechanisms. As a research method is also full of utilities and it has been mentioned in the book written by Catherine Belsey, Swansea University [2] in which it has been described that textual analysis as a research method is very important and why is it required; giving examples of an overview of a preliminary analysis of this one text which describes the images of women and the nature of rape, a very interesting take on textual analysis and why it could also be used to analyze such ancient overviews as well, to draw more insights. Applications in law are also imminent when textual analysis is concerned, a team of Rajiv Gandhi Institute of Technology researchers coupled this technology with other machine learning approaches to develop a system for predicting the result of judicial proceedings [7]. Another research work aims to use the knowledge discovery technique to a judicial judgments database in order to uncover the trend of opinion that Brazilian courts have in respect to the preferred party. As a result, their project aims to deliver software that enables law firms to get information in a rapid visual and exploratory manner, allowing them to focus and make more efforts in identifying more successful legal strategies rather than jurisprudential study [8]. Our research work also addresses a few components of textual analysis on lawful documents as well as political documents, history archives etc.

II. MOTIVATION

The aim of our project is to classify and analyze a certain dataset, perform NLP techniques to classify certain aspects of a text, specifically analysis of text for context based on input premise and hypothesis. We aim to have an accurate model which would predict certain anomalies and contradictions on the given input text. It's going to be a highly sophisticated NLP based project.

Applications of our project would range from identifying contradictions in documents used in the field of law to classifying relevant data from agenda documents (political, management etc.) – to draw conclusions from a set of text through textual pattern recognition.

Our aim with our research is to intensively explore the technologies like NLP and Transformers and based on these, work on textual analysis – contradiction detection and inconsistency identification.

III. BACKGROUND

Contextual analysis is basically the analysis of context from a particular dataset and then identifying the contradictions through a particular context and hypothesis; in addition to that, there could very well be inconsistencies in the text which could be identified as well. Two statements could be told apart through use of advanced NLP algorithms in terms of their context and relevance. Analysis of context and finding discrepancies has a lot of applications in the real world which makes it a great utility.

IV. Related Work

Here in this research work, Christopher D Manning, Marie Catherine de Marneffe and Anna N. Rafferty have studied about finding contradictions in text, where they explored different ways of contradictions that could occur through texts and datasets and work on a system that could automatically detect certain anomalies in those specific texts [3].

When watching discussions amongst democratic politicians, think about using a discrepancy identification program to help people go into the vast quantity of data provided by highlighting areas where politicians have divergent views. It's possible to use discrepancy identification in official documents to show what data requirements to just be verified more thoroughly.

They've thrown some clarity just on murky world of textual conflict in this article. They offer an NLP-friendly definition of inconsistency and a set of discrepancies. Using this information, they discover how discrepancy would be an uncommon occurrence which may be caused in a range of methods; they come up with a classification of discrepancy categories and compile statistics on various frequency.

In this paper, the authors focus on identification of linguistic reasoning occurrences between natural language text fragments, like as inconsistency, entailment, and positions. Specifically, authors examine the development of a Recognizing Textual Entailment (RTE) dataset used in viewer material validation [4].

By their understanding, this database is really the first repository of RTE with in social networking and confirmation domains predicated on organically arising disagreement in annotations statements in critical incidents addressed on Twitter [5].

In this paper, authors emphasize that there are many sites on the internet, each one authored by a different individual. These individuals often have divergent viewpoints, which results in wildly divergent layouts. It is critical for individuals to be able to assess the trustworthiness of something like the knowledge that consume when extracting this from the internet. Checking because not all resources contradict the material may help you decide what is and isn't reliable [6].

Previous attempts have been made to come up with a contradiction detection model. In 2021, a team of researchers tried multiple deep neural network techniques to find out the best fit for their database. Their main aim was to detect the contradictory claims in medical literature. Various techniques used were, BERT, bidirectional LSTM and GloVe. Out of the following models they concluded that BERT performs better than the rest [9]. This insight from the paper clearly eliminates two techniques for our research.

V. Proposed Methodology

According to our approach as shown in fig. 1 after gathering the information, we will divide the data into test data and training data. On the data we will perform feature engineering and analyze it further to get the best outcome. For easy classification we will seperate the data into 3 labels. Further we will break down the train data to create the best batch size for training. On our data we will apply different algorithms/methods like KNN, RandomForest, AdaBoost and XLM-RoBERTa. XLM-RoBERTa surpasses the BERT model and XLM in categorization, sequence labelling, and question answering, which is one of the reasons to use it [10].

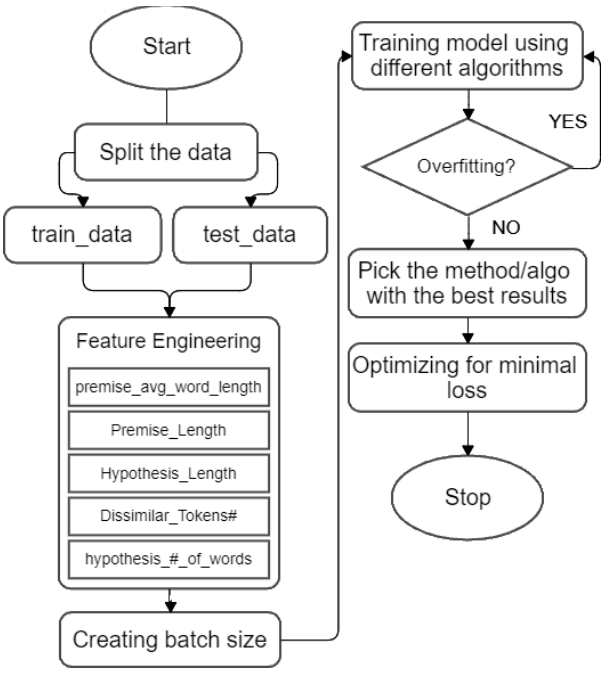

Fig. 1. Approach model

After checking for overfitting, we will choose the best method to further train our model. At the last step we processed the data and optimized the algorithm for minimal loss. We're primarily focusing on English as our language and going with a simple principle of using Premise and Hypothesis in two parts.

We're doing feature engineering in two sections:

i) One being the main features and data that we are extracting from the text that we are processing – being the number of words, the amount of punctuations, the amount of sentence clauses, the amount of sentence breaks and the amount of paragraphs as well to correctly assess the data we have to process.

ii) we're also considering the pure analytical part where the features that are based on text are considered including the quantities like the frequency, SVD, LSA, pervasive tools like word2vec and vec.

So, in the premise and hypothesis, we will observe the meta features of the dataset to firstly process the data:

The Meta features would involve the following factors: the number of words in the data, the unique words that are present in the data, the number of unique/underlying character in the text, macro analysis like checking for the number of small case letters and uppercase letters, the title case words that are there and also the combined average of the words that are present in the data. These features of the text allow our algorithm to process the data in a systematic way which we would then process in the NLP algorithm.

Here, we are analyzing a particular sentence and considering a few outcomes of it, we will decide based on the three that a sentence could either entail the other sentence,

could either contradict the sentence or both the sentences could be completely unrelated to each other.

So, we are creating a Natural Language Inference here which would assign a particular label to the sentences after determining the outcome of the sentence pair that is being evaluated and processed under the NLP. All of these are under a premise and hypothesis.

The specifics of our data are as follows, we have a dataset that contains a premise and a hypothesis. We will explore our data and read into the number of rows and columns in train data and test data.

Performing the code analysis of train_df = pd.read_csv("dataset") and test_df = pd.read_csv("dataset"), dividing the dataset into two parts. We will assign the variable x1, y1 to rows and columns of train data and assign the variables x2,y2 to the rows and columns of test data.

Now we are going to explore the three output labels 0,1,2: namely entailment, contradiction and neutral.

We have the following count after running a Data Frame:

0 – Entailment – 4176

1 – Contradiction – 4064

2 – Neutral – 3880

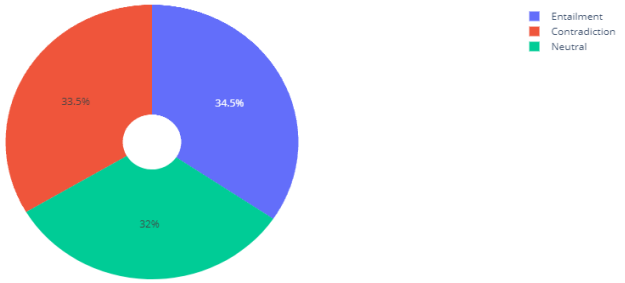

Fig. 2. Percentage distribution of the 3 classes.

So, from fig. 2 it can be concluded that we have the following observation made as per the total count of observations among the 3 labels in the NLP, being entailment, contradiction and neutral.

As the dataset is English intensive, we have tried to optimize this fact and convert it into an advantage. Constructing a percentage distribution through a dataframe, we realize that English is 56.7% of the data among different languages. We have in conclusion observed that the distribution between train and test is uniform.

VI. Result and Analysis

We will be taking in all the factors into consideration as features and will narrow it down under the train data and test_data. After that, we will continue and create a dictionary for the language column to enable reparsing of encoding if necessary.

1. We will then add the features and represent them into tokens and how they perform similarly in the premise and the hypothesis, initializing a tokenizer.

2. After initializing the tokenizer, we will pass them into the test and train data after which the data would be trained necessarily.

3. After the training is done, we would want to acquire and undertake the most common features as tokens which are present in each of the labels – entailment, contradiction and neutral. The following would give us insights into the data, and we would have a common token for every existing label that is.

4. After obtaining the common token we can approach this by creating a Boolean feature that passes us the info if the label contains the most common word or does it not which gives us the data to train on, hence after this, we train data using train_data.

Getting the output, we run into a few problems which would be solved by performing Exploratory Data Analysis: The data output we get gives us the mean length of the Premise and mean length of the hypothesis and their respective similar tokens.

Here, we are co-relating the values in the following way:

Premise – 107: 3.6
Hypothesis – 54: 3.17

This gives us the similarity index between the Premise and Hypothesis. Plotting the following graph in different languages and their count:

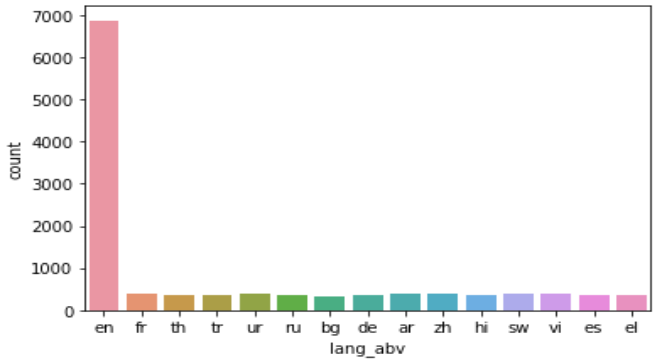

Fig. 3. The graph represents the languages present in the dataset with their corresponding count.

Now from the information obtained from fig. 3, we can clearly recognize that English is highly intense in the dataset and since our samples in the data are based mostly on English.

In the languages, English being the highest, we are considering the labels, accordingly, namely valued in 0,1,2:

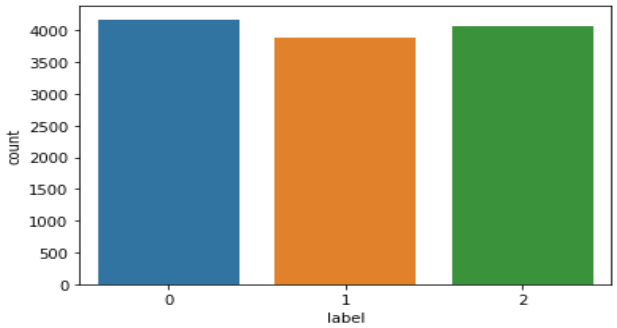

Fig. 4. The graph shows the count of each label in the dataset for the English language.

The chances of overfitting here is very less since we can see in fig. 4, our data has a label distribution which is very similar to each other, and hence overfitting to one certain label is a very minimal chance. We can further check the spread of Premise and Hypothesis lengths, average word length and the spread of similar and dissimilar tokens.

Here's the example of the spread of hypothesis length and dissimilar tokens in the data via labels:

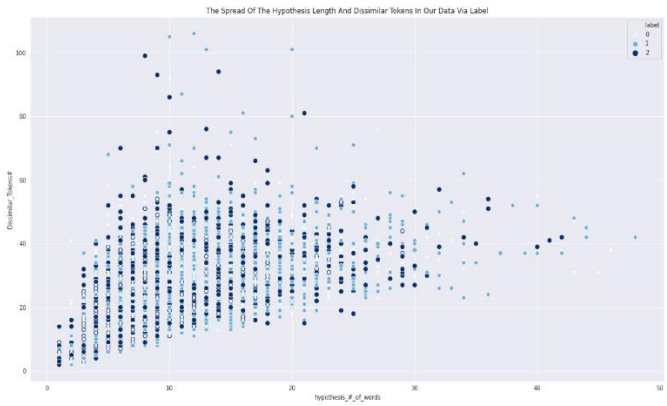

Fig. 5. Scatter plot for the hypothesis length and dissimilar tokens.

From fig. 5 it can be observed that the data is showing an overall uphill trend, hence it can be said that there is a positive relationship between hypothesis of words with dissimilar tokens. The next step to the process would be removing outliers from the dataset, performing winsorization.

Here, we will plot the heatmap now to draw correlations in the features that we have in the data:

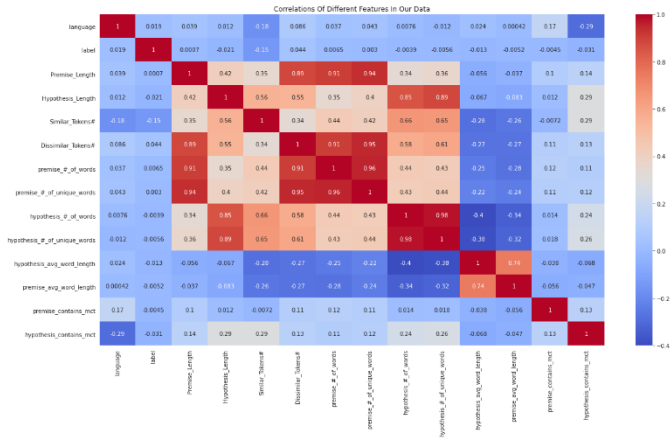

Fig. 6. Heatmap for correlations of different features

In fig. 6, we see that our ordinal data characteristics exhibit no association with the target labels, apart for those that have a negative association with related tokens.

After seeing that we do not have a very high quantitative connection amongst our features and our labels, any regression type methods are eliminated from consideration, leaving us with a more 'obvious' categorization technique to choose from instead.

VII. Comparison

We will go through a few models and test their accuracy according to the way the models perform, going through the models, we will see if they fit or not and if they don't work out as we prefer, then we will build a neural network as well.

Now, splitting the train data and trying the models one by one, we will analyze the outcomes.

Here we have the KNN model where we will check for the accuracy according to the number of neighbors that we have here. After testing, we observed that our model has the accuracy of around 0.435, as shown in fig. 7, when we plot the graph according to the N value.

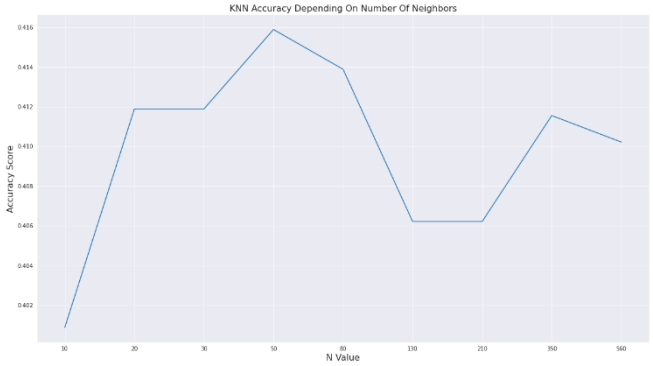

Fig. 7. KNN accuracy graph

Due to the unsatisfying results, we implemented and used the RandomForest Model. After running this model, we get the accuracy; RandomForest's accuracy depends on the number of estimators.

Comparing the accuracy of this model with N score in the graph:

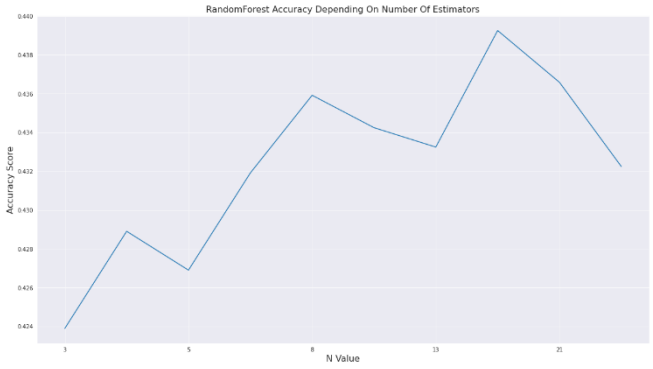

Fig. 8. RandomForest accuracy graph

From the fig. 8 we observed that the Random Forest model just did a little better than the KNN model, so we decided to move on to the next model.

We also tried the AdaBoost model:

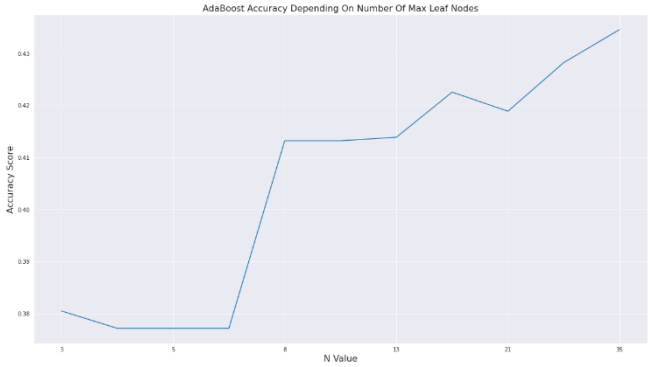

Fig. 9. AdaBoost accuracy graph

Here, we observed that this model is also giving us similar accuracy which is around 0.45, as can be seen in fig. 9, so to

exponentially increase the accuracy, we decided to try XLM RoBERTa, we could've also used a Neutral Network, but XLM Roberta worked out in our favour.

Using the XLM RoBERTa, importing and implementing the tokenizer and then training the model according to our 'hypotheses' and 'premise': Firstly, we converted the input ids into tensors and trained accordingly and finally implemented it.

The accuracy achieved by using XML RoBERTa was way better than the KNN and RandomForest models, so we decided to go with XLM RoBERTa where the accuracy reaches ~0.91. The model accuracy and loss can be observed from fig. 10 and fig. 11.

Here are the following outcomes in graphs:

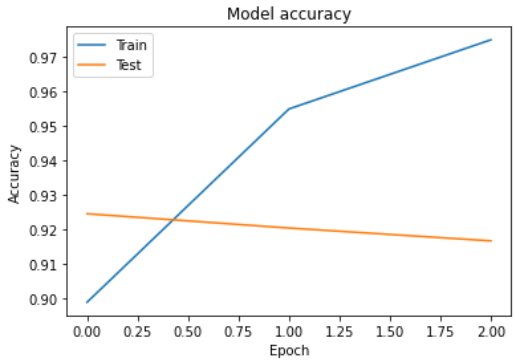

Fig. 10. XLM RoBERTa model accuracy graph.

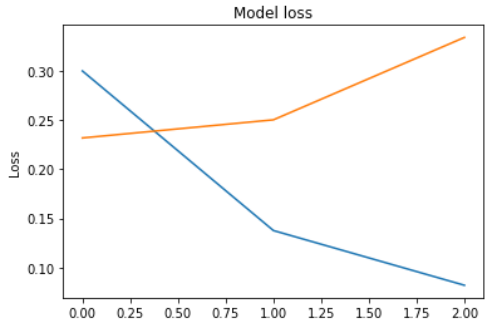

Fig. 11. XLM RoBERTa model loss graph.

Here, we get the output R2 score as: 0.902414927313117

VIII. Conclusion

So, in conclusion to our project we were able to devise the sentence pairs into 3 labels being --entailment, contradiction and neutral which classified the pairs on the basis of the data that we passed and made necessary deductions according to the NLP-based model training that we pass it through, we tried KNN, Random Forest, tried making Neural Networks and in the end settled up with using XLM RoBERTa which gave us the highest accuracy, as shown in fig. 10, and hence performed comparative analysis so that we can conclude and move with the final model which gave us the best results based on our features. With this, we aim to solve problems in fields of law, politics, and archaeological documents where documents require to be assessed manual observation, our model will classify them into 3 sections. Making them easier to assess.

Finally, we were able to predict our results with around ~90% accuracy in the dataset that we have used.

References

[1] Sanda Harabagiu and Andrew Hickl. Methods for using textual entailment in open-domain question answering. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 905–912, 2006.

[2] Catherine Belsey. *Textual analysis as a research method*. Edinburgh: Edinburgh University Press, 2013.

[3] Marie-Catherine De Marneffe, Anna N Rafferty, and Christopher D Manning. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, 2008.

[4] Piroska Lendvai, Isabelle Augenstein, Kalina Bontcheva, and Thierry Declerck. Monolingual social media datasets for detecting contradiction and entailment. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4602–4605, 2016.

[5] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. A survey on stance detection for mis-and disinformation identification. *arXiv preprint arXiv:2103.00242*, 2021.

[6] Rob Ennals, Dan Byler, John Mark Agosta, and Barbara Rosario. What is disputed on the web? In *Proceedings of the 4th workshop on Information credibility*, pages 67–74, 2010.

[7] Bhilare P, Parab N, Soni N, Thakur B. Predicting outcome of judicial cases and analysis using machine learning. Int Res J Eng Technol (IRJET). 2019;6:326-30.

[8] Barros R, Peres A, Lorenzi F, Krug Wives L, Hubert da Silva Jaccottet E. Case law analysis with machine learning in Brazilian court. InInternational Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems 2018 Jun 25 (pp. 857-868). Springer, Cham.

[9] F. S. Yazi, W. -T. Vong, V. Raman, P. H. H. Then and M. J. Lunia, "Towards Automated Detection of Contradictory Research Claims in Medical Literature Using Deep Learning Approach," 2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP), 2021, pp. 116-121, doi: 10.1109/CAMP51653.2021.9498061.

[10] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116. 2019 Nov 5.