Detecting Contradiction and Entailment in Multilingual Text

1st Abhigya Verma
Indira Gandhi Delhi Technical University for Women
Delhi, India

2nd Jahnvi Srivastav
Indira Gandhi Delhi Technical University for Women
Delhi, India

3rd Pooja Gera
Indira Gandhi Delhi Technical University for Women
Delhi, India

4th A. K. Mohapatra
Indira Gandhi Delhi Technical University for Women
Delhi, India

2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-3509-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCCNT56998.2023.10307714

***Abstract*—With many applications, including text categorization, question-and-answer systems, and sentiment mapping, entailment and contradiction detection in multilingual content is a crucial task in the discipline of natural language processing. This study examines the capabilities of two advanced models, the BERT-based multilingual case model and the XLM-RoBERTa large model, in detecting entailment and contradiction in a multilingual dataset containing 15 diverse languages. While the initial performance of BERT and XLM-RoBERTa models was 30% and 40% respectively, we have successfully enhanced the XLM-RoBERTa model's accuracy to an impressive 70% using data processing techniques. These results imply that, by carefully selecting models and employing the relevant data processing techniques, it is possible to detect contradiction and entailment in multilingual text with high accuracy. The study's outcomes are relevant to various areas of natural language processing, including the analysis of multilingual text. These results lay the groundwork for future studies in language processing and can guide the development of more sophisticated models.**

***Index Terms*—NLP, BERT, Multilingual Text, Classification**

I. Introduction

As the amount of multilingual content on the internet continues to grow, accurately identifying entailment and contradiction in this type of text have emerged as critical challenges. The internet has seen rapid growth in multilingual content in today's world, creating a pressing need for effective methods for detecting contradiction and entailment. It is essential in NLP applications to ensure proper understanding and interpretation of the text. Text classification, machine translation, question-answering systems, and sentiment analysis are some practical applications. This has recently been accomplished using language models which are pre-trained, such as BERT [1], RoBERTa [2], and XLNet [3], which have been demonstrated to be effective in cross-lingual settings.

This study employed the BERT-based Multilingual Cased model and XLM-RoBERTa large model to meet our objective to identify entailment and contradiction in the multilingual text. The data was preprocessed by translating non-English texts into English texts to improve the accuracy of XLM-RoBERTa, which achieved up to 70% accuracy. The study also utilized diverse data analysis techniques to investigate the dataset's class distribution, language diversity, and sequence lengths which helped us to refine our models accordingly.

The comparative examination of contradiction and entailment in multilingual literature has not been extensively explored in prior research work. This paper aims to fill this research gap by working on a novel dataset of multilingual data and facilitating an in-depth analysis across multiple languages. Our study distinguishes itself through the incorporation of a diverse range of models, namely BERT and RoBERTa, to attain a higher level of precision and depth in the detection of entailment and contradiction. In addition to greatly increasing the robustness of our findings, the incorporation of different encoding methods and data processing techniques increased the overall effectiveness of our investigation. The conclusions reached in this study have significant implications for the study of natural language in multilingual texts. By analyzing subtle differences in contradiction and entailment across numerous languages, our research provides valuable information and practical applications.

II. Literature Review

The detection of contradiction and entailment in multilingual text is a critical area of research in natural language processing. With the creation of potent language models like the BERT Multilingual base model(Cased) and XLM-RoBERTa, this field has progressed significantly. Through the use of techniques such as next-sentence prediction and masked language modeling, the BERT model has undergone pre-training on a vast range of 104 languages. It was first presented in the study titled [1] ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”.

On the other hand, XLM-RoBERTa is a scaled cross-lingual sentence encoder that has been trained on a vast dataset comprised of over 100 languages, which was carefully selected from Common Crawl. It was first presented in the study titled, [4] ”Unsupervised cross-lingual representation learning at scale” These two models have shown significant potential in detecting entailment and contradiction in multilingual text and have been widely used in various natural language processing tasks.

TABLE I
COMPARATIVE ANALYSIS BETWEEN PRIOR RESEARCH WORKS

Paper	Dataset Used	Task Performed	Model Used
[5]	Stanford Natural Language Inference (SNLI) dataset (110,000 sentence pairs) [6]	Identifying contradictions and entailment in a single language	Encoder-Decoder Sequence-To-Sequence Recurrent Neural Network
[7]	Machine-translated version of the Stanford Natural Language Inference (SNLI) corpus	Detecting contradiction and entailment in multilingual text	Recurrent Neural Network
[8]	A Hinglish SentiWordnet was created by combining the English and Hindi SentiWordnet	Detecting sarcasm and analysing sentiment in ”Hinglish” language	Extended sentiwordnet 3.0 and naïve Bayes classifier
[9]	Social Media Data	Sentiment analysis of social media text using a system to detect text and emoticons sarcasm	Artificial neural networks
[10]	Collection of 994 Text samples with sentiment and sarcasm labels along with eye-movement data from seven readers	Classification of Sentiments and Identifying Sarcasm	DNN
[11]	Social Media Data and News Headlines dataset	Sarcasm Detection	DNN
[12]	Retrieved Data from Amazon dataset	Sarcasm Detection and Sentiment Analysis	K Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random forest
[13]	Twitter Dataset and Internet Argument Corpus v2	Sarcasm Detection	BERT and Deep Learning Model

Majumder et al. [10] argue that sentiment classification and sarcasm identification, though considered separate tasks, are correlated, and propose a multitask learning-based framework that models this correlation using a deep neural network. The method proposed in this study achieved a 3-4% improvement over existing methods on a benchmark dataset.

Ajnadkar et al. [11] investigated the use of deep neural networks in detecting sarcasm in social media and news headlines and highlight the importance of such detection in improving sentimental analysis. With word embeddings, bidirectional LSTM, and convolutional networks, deep neural networks achieve 88% accuracy in detecting sarcasm.

Rao et al. [12] discuss sentiment analysis of text on social media, including the detection of sarcasm. The process involves dataset selection, preprocessing, feature extraction, and classification using algorithms like SVM and Random forest with evaluation based on accuracy.

Eke et al. [13] presented an approach to detect sarcasm using a context-based feature technique combined with deep learning model BERT and traditional machine learning. The technique addressed the limitations of existing models. The technique was tested on two Twitter datasets and achieved high precision rates of 98.5% and 98.0%, respectively, and 81.2% on the IAC-v2 dataset, demonstrating its superiority over existing approaches for sarcasm analysis. Mandal et al. [14] presented a novel approach utilizing a deep neural network architecture combining convolutional neural networks and Long Short-term Memory layers to detect sarcasm in news headlines with 86.16% accuracy. Sengar et al. [15] proposed an innovative method for identifying sarcasm in plain text by using feature engineering based on contrasting words within sarcastic sentences. The proposed method applies a ReLU activation function neural network model to improve the f1-score, while also capturing contextual data, in contrast to traditional machine learning techniques.

Raviraj Joshi. [16] introduces L3Cube-MahaCorpus, a Marathi monolingual corpus with 24.8M sentences and 289M tokens, along with BERT-based models (MahaBERT, MahaAlBERT, and MahaRoBERTa), a generative pretrained transformer model (MahaGPT), and Marathi fast text embeddings (MahaFT) trained on the full corpus.

Jallad et al. [17] address the challenging task of detecting contradictions in Arabic text by creating a dataset called ArNLI and proposing a novel approach that combines contradiction and language model vectors as input to a machine learning model. Promising results are achieved, with accuracies of 60%, 99%, and 75% on the SICK, PHEME, and ArNLI datasets, respectively, using a Random Forest classifier.

III. Methodology

The methodology section of our study outlines the approach we used to investigate the potential of advanced language models in detecting contradiction and entailment in multilingual text. The growing requirement for efficient natural language processing (NLP) models that manages the complexity of multilingual data served as the driving force behind our investigation. To achieve our research objectives, we leveraged the ”Contradictory, My Dear Watson” [1] dataset, which consists of pairs of premises and hypotheses in 15 different languages.

The dataset was analyzed using various data processing techniques and evaluated the accuracy of different language models, including BERT-based multilingual case models and XLM-RoBERTa large models. The data preprocessing and model training methodology, along with the experimental setup and evaluation criteria utilized to evaluate the model's performance, are described in this section.

A. Description of Dataset

The dataset ”Contradictory, My Dear Watson” has been leveraged as a crucial resource in our study. The dataset

[1] https://www.kaggle.com/competitions/contradictory-my-dear-watson

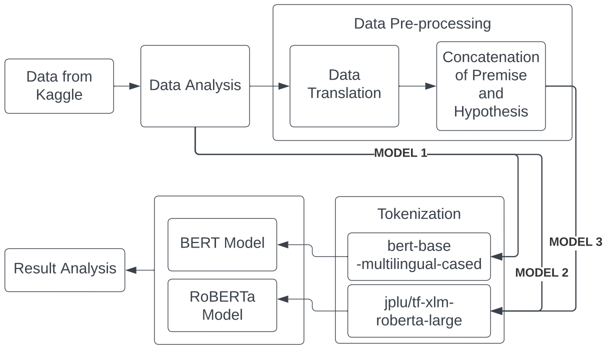

Fig. 1. Flow diagram representing our Methodology

contains pairs of premises and their hypothesis in diverse languages, making it suitable for a broader audience. The relationship between the premise and the hypothesis in each pair of the dataset is determined using the details provided in the premise.

The dataset comprises a testing set without labels as well as a labeled training set. The training set includes premise-hypothesis pairs, their ID, label, and language, while the testing set includes the same, except for labels. The dataset exploration allowed for the identification of critical variables and connections between them.

The analysis of the dataset properties revealed several key findings. The dataset contains 4176 rows with label 0, 4064 rows with label 2, and 3880 rows with label 1, with no duplicates identified. Additionally, the study disclosed the presence of 8,209 unique premises, 12,119 unique hypotheses, and 15 unique languages in the dataset. The identification of these properties and their relationships facilitates the development of robust models and data processing techniques that account for the dataset's intricacies.

B. Data analysis and Pre-processing

The class distribution analysis 2 revealed an equitable distribution of the dataset among the three classes, with class 0 exhibiting the largest share, trailed by class 2 and class 1.

In addition, the language distribution 3 analysis divulged that the majority of the examples in the dataset were in English, with other languages demonstrating comparable representation. Notably, as the definition of a word varies across languages, we opted to count the number of characters and tokens instead.

We performed a sequence length analysis to quantify the number of characters in both the 'premise' column as presented in 4. The results showed that the premises' length was relatively short, with the minimum and maximum being 4 and 967, respectively. The median length was 96. Furthermore, the sequence lengths were similar across all three classes, as demonstrated by the generated boxplot.

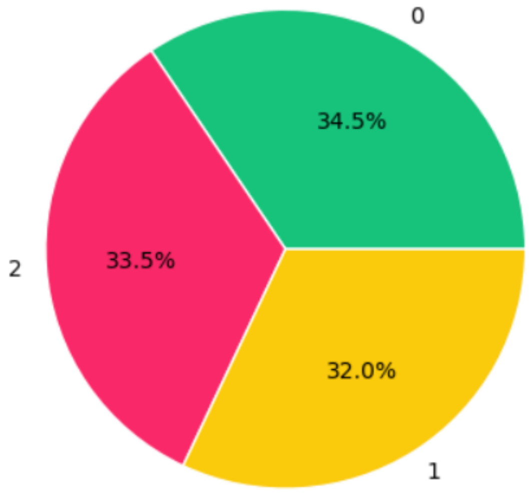

Fig. 2. Distribution of the Different Categorical Attribute: Class

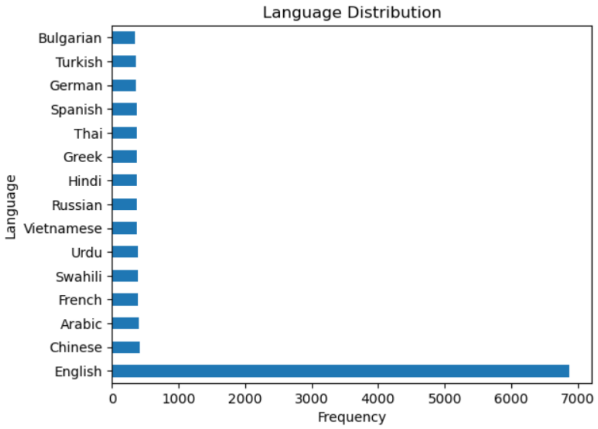

Fig. 3. Distribution of the Different Categorical Attribute: Language

Our data analysis uncovered a significant insight - the majority of premises were in English. To improve our accuracy, we employed a data preprocessing technique that involved creating a new data frame comprising non-English examples. The data was translated into English using the Google Translate library, and the robust RoBERTa model was then trained. This approach not only led to an impressive increase in accuracy but also highlights the crucial role of efficient data preprocessing in improving our present models.

C. Tokenization

Tokenization, the technique of splitting words into smaller parts known as tokens, is a crucial stage in the natural language processing process. Various pre-trained tokenizer models are available, including BERT and XLM-RoBERTa, which can be used to tokenize multilingual text. Following data analysis

TABLE II
DATASET ATTRIBUTES DESCRIPTION

Attribute	Description	Data Type
id	Unique Identification character sequence	String
premise	A sentence serving as the input text for the natural language inference task.	String
hypothesis	Represents a statement that is either entailed by, contradicts, or is neutral to the premise. It is the statement that we are trying to predict whether it can be inferred from the given premise or not, and what kind of relationship it has with the premise.	String
lang_abv	Abbreviation representing the language of the premise and hypothesis text.	String
language	The full name of the language used in the premise and hypothesis text.	String
label	The labels in this research paper represent the logical relationship between pairs of sentences and are assigned values of 0, 1, or 2 to indicate whether the relationship is Entailment, Contradiction, or Neutral.	Integer

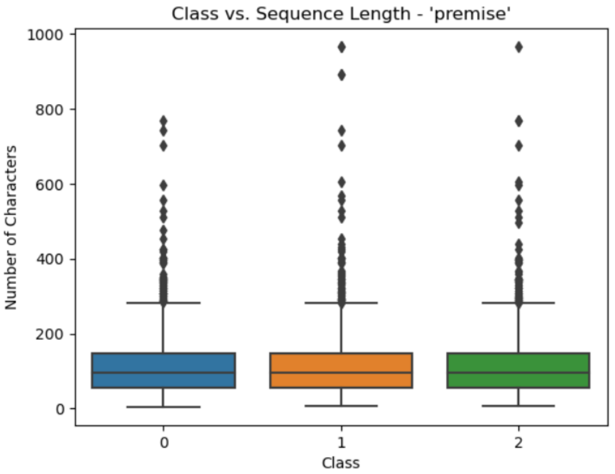

Fig. 4. A boxplot is generated to observe the variation of sequence lengths by class, which indicates that the sequence lengths are quite similar across the classes.

and preprocessing, we have utilized the bert-base-multilingual-cased and jplu/tf-xlm-roberta-large tokenizers in the BERT multilingual base and XLM-RoBERTa models, respectively.

D. Model Construction

1) BERT multilingual base model (cased):: The BERT model construction includes three input layers: input_word_ids, input_mask, and input_type_ids. The TFBertModel layer comprises the BERT model architecture, with a total of 177,853,440 parameters. This layer outputs a TFBaseModelOutputWithPoolingAndCrossAttentions object, including the last hidden state with shape (None, None, 768), pooler output with shape (None, 768), and other attributes such as past hidden states, attentions, key values, and cross-attentions. The tf.__operators__.getitem layer slices the last hidden state to output a tensor with shape (None, 768). Finally, the dense layer is used for classification with 3 output classes. The total trainable parameters of the BERT model are 177,855,747.

2) XLM-RoBERTa (large sized model):: The construction of XLM-RoBERTa model includes one input layer, input_layer, with shape (None, 120). The tfxlm_roberta_model_1 layer comprises the XLM-RoBERTa model architecture, with a total of 559,890,432 parameters. This layer outputs a TFBaseModelOutputWithPoolingAndCrossAttentions object, including the last hidden state with shape (None, 120, 1024), pooler output with shape (None, 1024), and other attributes such as hidden states, attentions, past key values, and cross-attentions. The tf.__operators__.getitem_1 layer slices the last hidden state to output a tensor with shape (None, 1024). The dropout_148 layer applies dropout regularization to the output of the slice layer. The subsequent dense layers dense_5, dense_6, dense_7, dense_8, and dense_9 have output sizes of 64, 32, 16, 8, and 3, respectively, and are used for classification. The total trainable parameters of the XLM-RoBERTa model are 559,958,803.

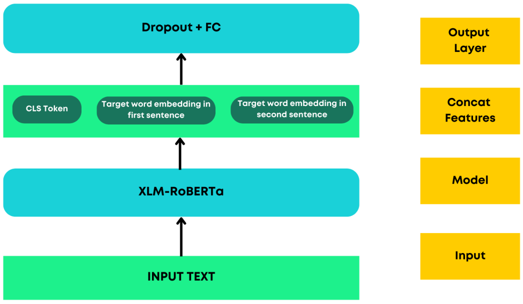
Dropout + FC
CLS Token
Target word embedding in first sentence
Target word embedding in second sentence
XLM-RoBERTa
INPUT TEXT
Output Layer
Concat Features
Model
Input

Fig. 5. XLM-RoBERTa model architecture

IV. RESULTS AND DISCUSSION

The current study aims to detect contradiction and entailment in multilingual text using different models. The initial approach utilized the BERT model, which achieved an accuracy of 30%. The subsequent approach involved the RoBERTa model, which improved the accuracy to 40%. However, by employing data processing techniques such as converting non-English text to English text, the accuracy was significantly enhanced to 70%. The results provide insight into the potential of machine learning models such as BERT and RoBERTa to detect complex linguistic relationships in multilingual text. The study also highlights the significance of employing appropriate data processing techniques to improve the accuracy of natural language processing models. These findings have significant implications for developing advanced

TABLE III
COMPARITIVE ACCURACY FOR THE APPLIED MODELS

Model Used	Tokenizer Used	Accuracy
Bert	bert-base-multilingual-cased	30%
Roberta	jplu/tf-xlm-roberta-large	40%
Roberta with Data Processing	jplu/tf-xlm-roberta-large	70%

natural language processing models capable of accurately identifying entailment and contradiction in multilingual text.

V. CONCLUSION AND FUTURE SCOPE

This research has demonstrated the potential for identifying and understanding entailment and contradiction in multilingual text by leveraging advanced language models such as the Bert-based multilingual case model and XLM-Roberta large model. With an impressive accuracy of 30% and 40% respectively, the application of data processing techniques to translate non-English languages into English has paved the way for significant improvements in accuracy, with the XLM-Roberta model achieving up to 70%. The study underlines the significance of data processing techniques, and further investigation is necessary to explore more efficient methods for detecting contradiction and entailment in multilingual text, as well as developing models that can handle language variations and nuances more effectively. Future research may employ the latest advanced models and focus on training models on proprietary data prior to application. This study has opened up new possibilities for NLP research, emphasizing the importance of developing robust models for multilingual applications.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[3] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

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

[5] R. Sifa, M. Pielka, R. Ramamurthy, A. Ladi, L. Hillebrand, and C. Bauckhage, "Towards contradiction detection in german: a translation-driven approach," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2019, pp. 2497–2505.

[6] M. Glockner, V. Shwartz, and Y. Goldberg, "Breaking nli systems with sentences that require simple lexical inferences," *arXiv preprint arXiv:1805.02266*.

[7] M. Pielka, R. Sifa, L. P. Hillebrand, D. Biesner, R. Ramamurthy, A. Ladi, and C. Bauckhage, "Tackling contradiction detection in german using machine translation and end-to-end recurrent neural networks," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 6696–6701.

[8] A. Gupta, A. Mishra, and U. S. Reddy, "Sentiment analysis of hinglish text and sarcasm detection," in *Conference Proceedings of ICDLAIR2019*. Springer, 2021, pp. 11–20.

[9] S. Gupta, R. Singh, and V. Singla, "Emoticon and text sarcasm detection in sentiment analysis," in *First International Conference on Sustainable Technologies for Computational Intelligence: Proceedings of ICTSCI 2019*. Springer, 2020, pp. 1–10.

[10] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, and A. Gelbukh, "Sentiment and sarcasm classification with multitask learning," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 38–43, 2019.

[11] O. Ajnadkar, "Sarcasm detection of media text using deep neural networks," in *Computational Intelligence and Machine Learning: Proceedings of the 7th International Conference on Advanced Computing, Networking, and Informatics (ICACNI 2019)*. Springer, 2021, pp. 49–58.

[12] M. V. Rao and C. Sindhu, "Detection of sarcasm on amazon product reviews using machine learning algorithms under sentiment analysis," in *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. IEEE, 2021, pp. 196–199.

[13] C. I. Eke, A. A. Norman, and L. Shuib, "Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and bert model," *IEEE Access*, vol. 9, pp. 48 501–48 518, 2021.

[14] P. K. Mandal and R. Mahto, "Deep cnn-lstm with word embeddings for news headline sarcasm detection," in *16th International Conference on Information Technology-New Generations (ITNG 2019)*. Springer, 2019, pp. 495–498.

[15] C. P. S. Sengar and S. Jaya Nirmala, "Sarcasm detection in tweets as contrast sentiment in words using machine learning and deep learning approaches," in *Machine Learning, Image Processing, Network Security and Data Sciences: Second International Conference, MIND 2020, Silchar, India, July 30-31, 2020, Proceedings, Part I 2*. Springer, 2020, pp. 73–84.

[16] R. Joshi, "L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources," *arXiv preprint arXiv:2202.01159*, 2022.

[17] K. A. Jallad and N. Ghneim, "Arnli: Arabic natural language inference for entailment and contradiction detection," *arXiv preprint arXiv:2209.13953*, 2022.