ISSN No:-2456-2165

Sentence Prediction using BERT based on Malayalam

Arya K A[1] and Shelja Jose M[2] *
1 Second year MSc Computer Science, Vimala College (Autonomous), Thrissur
2 Assistant. Professor, Department of Computer Science Department, Vimala College (Autonomous) Thrissur

Abstract:- This paper deals with advancements that have occurred in the field of Natural language processing. Most of the text prediction-related tasks are conducted in different dialects but not in Malayalam. Numerous frameworks with various techniques were produced for various dialects. Only a few of the models fit into Malayalam. In this paper, the BERT algorithm is applied to the Malayalam language. BERT uses bi-directional pre-training on multiple tasks such as Next Sentence Prediction and Masked Language Model. This paper mainly deals with the Next sentence prediction.

Keywords:- *BERT, Malayalam Next Sentence Prediction.*

I. INTRODUCTION

A lot of advancement has occurred in the field of natural language processing during the last few years. But for regional languages like Malayalam, needs more advancements and Implementations. The Indian subcontinent itself has a variety of languages, dialects, and styles spoken by more than a billion people. The iNLTK libraries provided the languages Hindi, Punjabi, Sanskrit, Gujarati, Kannada, Malayalam, Nepali, Odia, Marathi, Bengali, Tamil, and Urdu. Here the model is trained using Natural Language Toolkit for Indian Languages (iNLTK). The data set consists of 12k Wikipedia Articles that have been cleaned. It consists of around 6300 news article headlines. A corpus is a collection of headlines tagged with their news category. The paper mainly performs Next sentence prediction.

II. METHODOLOGY

The advancement of some fantastic transfer learning approaches like ELMO, ULMFIT, OpenAI, GPT, and BERT has seen the rise of NLP during the last two years. These algorithms are trained as language models on a large number of unstructured content corpora and then used for a variety of downstream NLP tasks like text categorization, question answering, paraphrase detection, and missing word prediction, among others.

The application of Transformer bidirectional training is BERT's key technical breakthrough. Previously, researchers looked at a text sequence from left to right or a combination of left-to-right and right-to-left training. One of the main use of BERT is in the next sentence prediction.

A Dravidian language like Malayalam has a very difficult grammar structure compared to other languages. Pragmatics is the sub-field of computational linguistics and also the sub-field of semiotics.

BERT (Bidirectional Encoder Representations from Transformers) has a transformer model (encoder-decoder architecture) pretrained on Wikipedia data from 104 languages including Malayalam. It is trained for two tasks: Masked language modelling and Next sentence prediction.

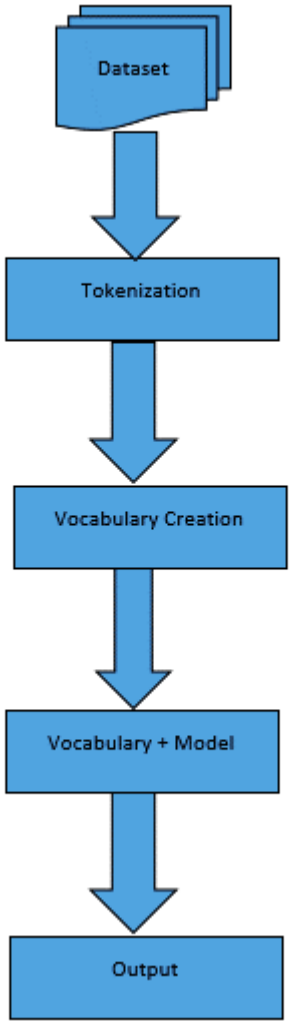

Fig 1: A conceptual overview of the system

A. Model Architecture

After collecting the dataset the first task to perform is tokenization. Taking every word in the model and then converted into a token. The token generated here is used for vocabulary creation. The Malayalam Tokenizer is available with the iNLTK library [1]. The iNLTK library is used for various applications like Tokenization, word embedding, text completion, the similarity of sentences, etc.

B. Training

For training first load the training data, testing data, and validation data. And also the vocabulary. After that, the batch for training is created. And the size of training is specified. Its value should be a multiple of eight. After that, the pre-trained model is loaded. It uses LSTM by default [2].

C. Dataset

There are two datasets is used for the prediction. The first is Wikipedia articles in Malayalam. This data collection includes sanitized Wikipedia articles of a total of 12,000. It has an 8.6k article in the Train set and a 3.7k article in the Validation set, that were used to train and assess Language Models for Malayalam in the repository NLP.

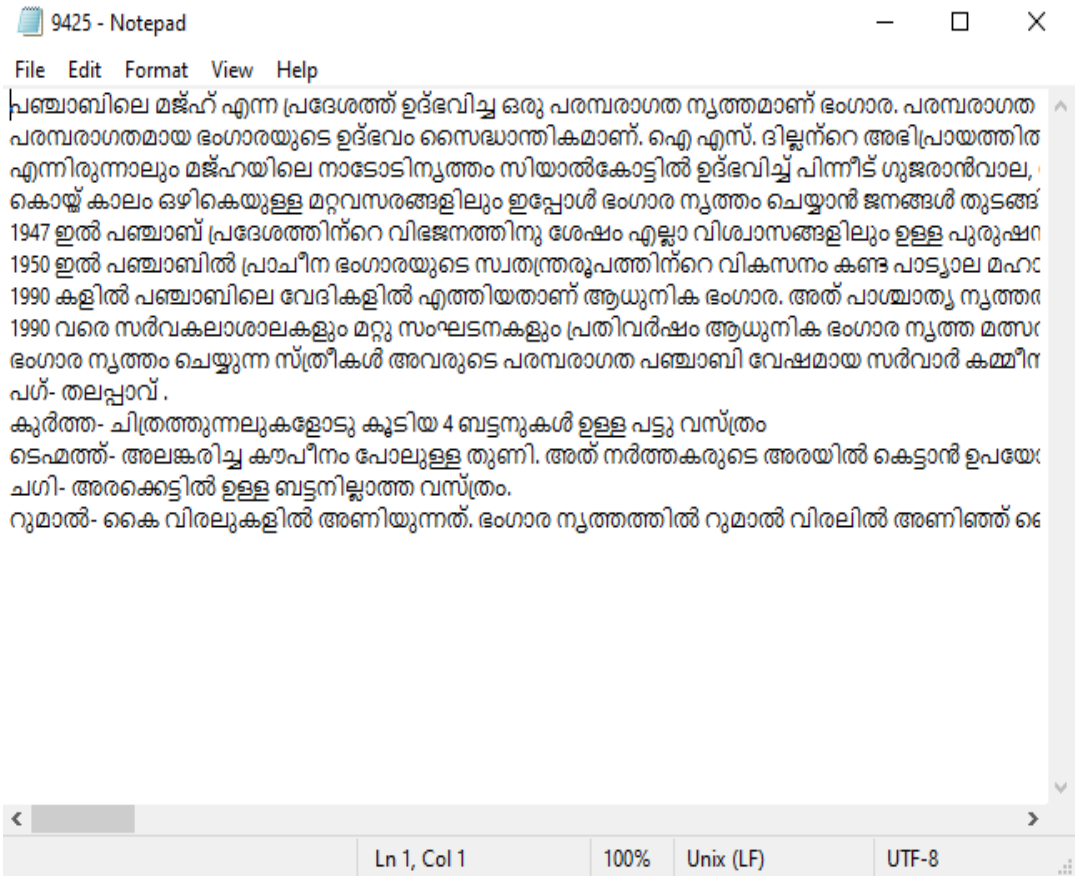

Fig: 2 Dataset

The next one is the Malayalam news dataset. This data set contains around 6300 news article headlines that had been collected from Malayalam news websites. The data set has been cleaned and contains train and test set data.

III. RELATED WORK

In every natural language processing application, the morphological analyzer has an important role. The issues in its development include both technical and language-dependent aspects. The different approaches for Malayalam morphological analyzers include the Hybrid approach, Paradigm Approach, and Suffix Stripping Approach. In the paradigm approach for building a morphological analyzer, it is necessary to create a morphological dictionary for improving accuracy. The suffix stripping module and post-processing module are introduced in the Suffix Stripping Approach. A lot of issues will face in the development of the Malayalam morphological analyzer.

Some of them are Lexical Entry and their Classification, Identification of proper paradigm for a word, Multiple suffix problem, Combining with the adjacent words, Occurrence of the same verb in different forms, Handling Proper nouns and Abbreviations, Identification of chunks, Handling of unknown words. If we don't have a Morph analyzer with considerably good accuracy then the accuracy of the whole system will get affected. When we consider the Machine Translation System the pattern identification and phrase structure transfer to the target language will be based on the Morphological Analyser output [3]. All the above-mentioned issues will be specifically applicable to the Dravidian languages and other agglutinative languages.

Natural language processing is one of the major fields in computer science. NLP is the ability of the system to process different sentences in natural language [4]. Parts of speech tagging, Pragmatic analysis, machine translation, discourse analysis, etc are the different fields in Natural language processing. It has a difficult grammar structure. The major takeaway from a pragmatics examination of Malayalam sentences is how sentences are employed in various situations or contexts. Pragmatics analysis is also known as speech act analysis.

It mainly considers how sentences are used in different situations. Most of the time, a sentence's effect is obviously declared, but in other instances, the intended effect is not. In these systems given the Malayalam sentence as the input and the output will be the classification of that particular sentence into different classes. The system has mainly four-module. They are Tokenization, Parts of speech tagging, Classifier, and Dynamic Corpus. Here the proposed system analyses the Malayalam sentence and classifies that Malayalam sentence into different classes. The main problem of the Malayalam works will be the unavailability of good quality Malayalam corpus and also the less number of Malayalam words in the corps is another major problem.

In many applications of natural language processing, natural language inference (NLI) is a crucial subtask. From the premise to the hypothesis, there is a directed relationship. The NLI is also known as textual entailment recognition, and it recognizes entailed and contradictory sentences in various NLP systems like Question Answering, Summarization, and Information retrieval systems [5]. Many datasets are available in the English language, and few for Italian, Japanese, and other languages. However, no datasets are available for the Malayalam language for entailment recognition. They are FraCaS, RTE datasets, SNLI, MultiNLI, XNLI, ANLI, SciTail, SICK, EVALITA, ArbTE, RITE dataset.

It compares different approaches Doc2Vec, Fast Text, mBERT, and LASER, to conclude that LASER embedding improved the system performance noticeably, emphasizing that improved input representations without much loss in information helped in obtaining better results [5]. The system performance is compared with XNLI results. The encoder models publicly available are utilized in obtaining sentence representations. Being a language-agnostic model, it can be extended to other languages also.

Word embedding's used to convert human language into a numerical form by encoding the semantic properties of words. Using it each word can be transformed into a set of N-dimensional vectors. It plays a vital role in the processing of linguistic applications like natural language inference, information retrieval, sentiment analysis, etc. The goal of word embedding is to capture the meaning of words in their context [2]. And it also finds the semantic relationships and similarities between words.

It discusses the different word embedding including Word2Vec, GloVe, fast Text, BERT, ELMo, XLNet, MUSE, IndicNLPSuite. Word embedding's becoming an essential part of processing linguistic applications. The models have not yet been fully created for languages with limited resources, including Malayalam. The absence of a large enough corpus is these languages' principal drawback. For such languages, fully functional word embedding's have not yet been created.

The prediction task in Natural Language Processing intends to figure out the missing characters, letters, words, expressions, or sentences. Numerous frameworks with various techniques were produced for various dialects. The missing content prediction is one of the significant concerns of Natural Language Processing. Moreover, most of the text prediction-related tasks are conducted in different dialects but not in Malayalam. This is because there are only a few standard models that could fit into Malayalam. The BERT model is based on the concept of isolating the pre-training and fine-tuning, which have already shown drastic improvements in the NLP [1].

No	Name	Author	Model
1	Natural Language Processing Based Text Imputation For Malayalam Corpora	Annlin Rojan, Edwin Alias, Georgy M. Rajan, Jithin Mathew , Dhanya Sudarsan	BERT
2	Pre-Trained Word Embedding's For Malayalam Language	Reji Rahmath K, P C Reghu Raj, Rafeeque P C	XLNet, GloVe, Word2Vec, fastText, MUSE,BERT, ELMo
3	Natural Language Inference For Malayalam Language Using Language Agnostic Sentence Representation	Sara Renjit and Sumam Idicula	LASER Doc2Vec fastText BERT
4	Pragmatic Analysis Of Malayalam Sentences	Shaharban T A, Rosna P Haroon	pragmatics analysis o
5	Issues In Development Of Malayalam Morphological Analyzer	Vinod P M, Jayan V, Bhadran V K	Suffix Stripping Approach, Paradigm Approach, Hybrid Approach

Fig: 3 Related Works

IV. RESULT AND DISCUSSION

The model proposed is capable of solving the general task of Next Sentence Prediction. The model has shown an accuracy of 58% for Next Sentence Prediction. The entire development and training of this model are done with the help of google colab and free Compute Engine of CPU. The entire data was stored in google drive and accessed remotely via the service of colab. Next sentence prediction output is as follows.

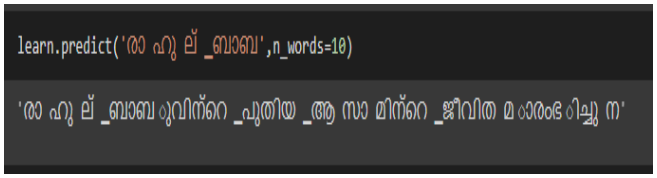

Fig 4: Result

V. CONCLUSION

In this work, the next sentence prediction for the Malayalam language is predicted by using the BERT algorithm. The model is capable of predicting the next sentence with an accuracy of 58%. With the use of more accurate data, the accuracy of prediction can be improved. Even though the model faces some constrained related to the Malayalam language.

The model proposed is capable of solving general tasks such as Next Sentence Prediction, Masked Language Modelling, QA, etc. This can also infer the knowledge from the Malayalam test corpus for widespread NLP problems. There are many packages are available for tokenization. None of these packages has built-in support for Indic languages, particularly Malayalam. If a more accurate Malayalam tokenizer is created then the model can be more effective.

ACKNOWLEDGMENT

We express our sincere gratitude to God Almighty for showering us with all blessings and express our gratitude to all the teaching staff for their valuable guidance and support at each stage of the work. We are also thankful to our parents for the support given in connection with the work.

REFERENCES

[1]. A. Rojan, E. Alias, G. M. Rajan, J. Mathew, and D. Sudarsan, "Natural language processing based text imputation for malayalam corpora," in 2020 *International Conference on Electronics and Sustainable Communication Systems* (ICESC), pp. 161–165, IEEE, 2020.

[2]. K. R. Rahmath, P. R. Raj, and P. Rafeeque, "Pre-trained word embeddings for malayalam language: A review," in 2021 *International Conference on Artificial Intelligence and Smart Systems* (ICAIS), pp. 568–572, IEEE, 2021.

[3]. P. Vinod, V. Jayan, and V. Bhadran, "Issues in development of malayalam morphological analyzer," in *2012 Annual IEEE India Conference* (INDICON), pp. 540–544, IEEE, 2012.

[4]. T. Shaharban and R. P. Haroon, "Pragmatic analysis of malayalam sentences," in 2016 *International Conference on Inventive Computation Technologies* (ICICT), vol. 3, pp. 1–5, IEEE, 2016.

[5]. S. Renjit and S. Idicula, "Natural language inference for malayalam language using language agnostic sentence representation*," PeerJ Computer Science*, vol. 7, p. e508, 2021.

[6]. 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.

[7]. Y. Liu, M. Ott, and N. Goyal, "Jingfei du, mandar joshi, danqi chen, omer levy, mike lewis, luke zettlemoyer, and veselin stoyanov. 2019. roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv*:1907.11692, vol. 32, p. 113, 2019.

[8]. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[9]. G. Chierchia and S. McConnell-Ginet, "Meaning and grammar: An introduction to semantics," 2000.