Contextual Embeddings Generation Using Custom XLM-RoBERTa Model (Team-23)

Abstract

In this report, we present a methodology for generating contextual embeddings using a custom XLM-RoBERTa model designed for extracting contextual embeddings. We also perform a Classification task as a testament of the quality of embeddings. We describe the custom model architecture, detail the embedding extraction process, and explain the rationale behind our chosen approach. This method offers a powerful way to generate meaningful and context-aware representations of words for Indic languages, resulting in improved performance over traditional word embeddings.

Introduction

Text classification is a fundamental task in Natural Language Processing (NLP) that involves assigning predefined categories or labels to text based on its content. Examples of text classification tasks include sentiment analysis, topic categorization, and spam detection. One of the key challenges in text classification is representing words in a way that captures their meaning in context. Traditional word embeddings, such as Word2Vec and GloVe, generate static representations that do not take into account the context of words within a sentence. Recent advancements in NLP, particularly the emergence of pre-trained transformer models like XLM-RoBERTa, have made it possible to generate contextual embeddings that greatly enhance the performance of text classification tasks.

Methodology

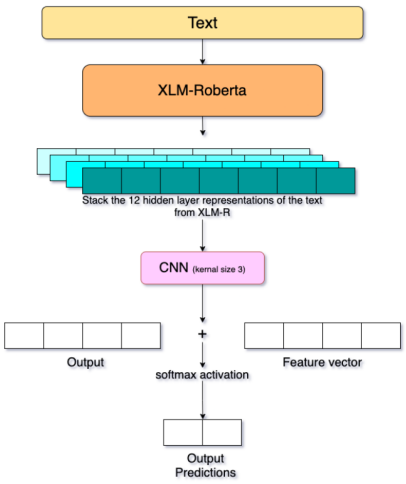
We propose a custom XLM-RoBERTa model to generate contextual embeddings. Our model architecture consists of the following components:

1. Pre-trained XLM-RoBERTa model: We leverage the XLM-RoBERTa base model as a starting point, as it has been pre-trained on large-scale multilingual data and can generate rich contextual representations of words.
2. Convolutional layer: A 1D convolutional layer is applied on the hidden states of the XLM-RoBERTa model. This layer captures local dependencies in the input sequence and helps in extracting meaningful features from the hidden states.

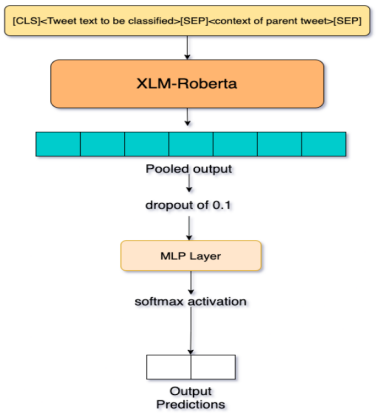
3. Max-pooling: We apply max-pooling on the output of the convolutional layer to retain the most important features and reduce the dimensions of the hidden states.
4. Dropout and fully-connected layer: We use dropout to prevent overfitting and a fully-connected layer to map the pooled features to the desired number of NER classes.

To generate contextual embeddings for a given input text, we follow these steps:

1. Tokenize the input text using the XLM-RoBERTa tokenizer, which is designed to handle indic(Devnagri script) text.
2. Forward the tokenized input through our custom XLM-RoBERTa model.
3. Extract the hidden states from the XLM-RoBERTa layers, which serve as the contextual embeddings for each token in the input text.



The figure alongside shows the base architecture for Hindi & Marathi languages for Subtask involving *Identifying Hate, offensive and profane content from the post* using XML-R with CNN augmented with textual features vector followed by a softmax layer.



This figure shows the model pipeline for hate detection in conversational threads for subtasks involving *Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL).*

Results

Our custom XLM-RoBERTa model achieved the following results for hate speech classification:

- Training accuracy: 95.37%
- Training precision: 95.29%
- Training recall: 94.59%
- Training F1 score: 94.92%
- Validation accuracy: 89.89%
- Validation precision: 86.95%
- Validation recall: 86.17%
- Validation F1 score: 87.53%

The best validation accuracy obtained during the experiments was 89.89%. These results demonstrate the effectiveness of our approach in classifying hate speech, with high accuracy and F1 scores indicating a good balance between precision and recall.

Language	***Method***	***Accuracy***	***Macro F1***
Marathi	XLM-R Base	84.16	0.86
	XLM-R + CNN	**88.64**	**0.87**

Discussion

Our proposed methodology for generating contextual embeddings using a custom XLM-RoBERTa model offers several advantages:

1. Context-aware representations: By leveraging the pre-trained XLM-RoBERTa model, our approach can generate embeddings that capture the meaning of words in context, leading to improved performance in classification tasks compared to traditional word embeddings.

2. Multilingual support: The XLM-RoBERTa model is pre-trained on a large-scale multilingual corpus, allowing our approach to support tasks in multiple languages.
3. Feature extraction: The use of convolutional and max-pooling layers in our model helps extract meaningful features from the hidden states, which can aid in better classification of named entities.
4. Regularisation: The inclusion of dropout helps prevent overfitting and ensures that our model generalises well to unseen data.

Conclusion

In this report, we have presented a methodology for generating contextual embeddings using a custom XLM-RoBERTa model for Named Entity Recognition tasks. Our approach leverages the power of pre-trained transformer models to generate context-aware word representations and incorporates additional layers for feature extraction and regularisation. The resulting model is capable of handling multilingual NER tasks and offers improved performance compared to traditional word embeddings.

Literature Review

- *A Comparative Study on Language Models for Dravidian Languages* by Rahul Raman, Danish Mohammed Ebadulla, Hridhay Kiran Shetty & Mamatha H.R. This paper trains embeddings for four Dravidian languages using the latest deep learning language models and evaluates their performance on text classification and small custom similarity tasks.
- *'A Passage to India': Pre-trained Word Embeddings for Indian Languages* by Kumar Sauravy, Kumar Saunack, Diptesh Kanojiay, and Pushpak Bhattacharyyay. This paper uses various existing approaches to create multiple word embeddings for 14 Indian languages and releases pre-trained embeddings generated using both contextual and non-contextual approaches2.

Additional Resources:

1. Presentation : Link
2. Model weights: Link