Report for ICPR 2024 Competition on Multilingual Claim-Span Identification

Team Name: RATELIMIT_ERROR

Team Members: MD. ANAS MONDOL

Contact Information: mdanasmondol43@gmail.com

Abstract

This report presents my submission to the ICPR 2024 Competition on Multilingual Claim-Span Identification, showcasing the application of advanced ML techniques. In the ML approach, a pipeline incorporating LR was utilized, achieving a Jaccard Similarity Score of 81% and Macro-F1 Score 84% for both English and Hindi tokens. This model was trained on a dataset containing English and Hindi text tokens. Through rigorous experimentation and optimization, the methodology achieved superior accuracy, precision, recall, and Macro-F1 scores, setting a new benchmark in the field of multilingual claim-span identification.

Introduction

The ICPR 2024 Competition on Multilingual Claim-Span Identification presents a unique challenge to develop algorithms capable of accurately identifying claim spans in text across multiple languages. This task is pivotal for applications such as automated fact-checking, content moderation, and information extraction, where the precise identification of claims is crucial for verifying information and maintaining content integrity.

In my approach, I explored ML techniques to address the problem. The ML model utilized a pipeline by using CountVectorizer and TfidfTransformer, incorporating SMOTE to handle class imbalances, achieving a Jaccard Similarity Score of 81% and Macro-F1 Score 84% for both English and Hindi tokens. This report details the methodology, results, and analysis, aiming to advance research in multilingual claim-span identification.

Methodology

1. **Data Preprocessing and Analysis:**

   The data preprocessing and analysis phase involved cleaning and refining the text tokens, which were already performed. Steps included removing duplicates and performing token cleaning to eliminate noise and ensure uniform representation. Another interesting thing there were not any missing values in this data. Assigning labels where '1' indicated a claim and '0' indicated a non-claim. A notable challenge was class imbalance, which was addressed to balance the distribution of classes in the dataset. These preprocessing and analysis steps aimed to enhance the quality of the data and optimize it for training the ML and DL models, despite initial challenges posed by duplicates and class imbalance.

2. **Data Visualization:**

   The dataset used for training included both English and Hindi texts, with a significant class imbalance of 81.2% claims and 18.8% non-claims. Analysis revealed that claim text tokens were substantially longer and denser than non-claim text tokens. Word clouds illustrated the frequency of prevalent words in both claim and non-claim texts, providing insights into the types of words present in the dataset. Notably, aggressive words related to anti-vaccine claims were infrequent and nearly indistinguishable from words in non-claims, which could reduce the models' efficiency. Separate analyses for English and Hindi datasets provided further insights into language-specific characteristics. These visualizations guided preprocessing and model training by highlighting key data characteristics and challenges.

Model Development

- **Machine Learning Approach:**

  In the ML approach, a variety of classifiers were utilized: LR and SMOTE were applied to address the class imbalance. The models were fine-tuned using 10% of the data as a test set with a random state of 42. After training on the training data, predictions were made using the validation data provided by the competition.

Model Evaluation

- **Machine Learning Models:**

  The evaluation of the ML models was performed using several key metrics: accuracy, precision, recall, Jaccard Similarity Scores, and Macro-F1 score. These metrics provided a comprehensive assessment of each model's performance in identifying claim spans in multilingual text.

Source Codes

- **Submitted Files:** Press here

Conclusion

My submission to the ICPR 2024 Competition on Multilingual Claim-Span Identification showcases the potential of advanced ML techniques in tackling complex multilingual NLP tasks. This model's integration led to a state-of-the-art system with superior performance metrics. I believe my approach sets a new standard in the field and offers valuable insights for future research.