Mini Project Report

on

Claim Span Identification using Deep Learning

Submitted by

Doddi Bhuvanesh,P Vamsi Madhav,R Dhivya Prasanna,Ashish Joy

21BDS018,21BDS051,21BDS052,21BCS016

Under the guidance of

Dr Sunil Saumya

Associate Dean, Academics

Dharwad
ज्ञानेन विकासः

INDIAN INSTITUTE OF
INFORMATION
TECHNOLOGY

DEPARTMENT OF DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DHARWAD

13/11/2024

Certificate

This is to certify that the project, entitled Claim Span Identification using Deep Learning , is a bonafide record of the Mini Project coursework presented by the students whose names are given below during 2024 in partial fulfilment of the requirements of the degree of Bachelor of Technology in Computer Science and Engineering.

Roll No	Names of Students
21BDS018	Doddi bhuvanesh
21BDS051	P Vamsi Madhav
21BDS052	R Dhivya Prasanna
21BCS016	Ashish Joy

$Dr.SunilSaumya$
(Project Supervisor)

Contents

List of Figures

List of Tables

1 Introduction

The proliferation of social media has revolutionized global communication, allowing individuals to connect and share information instantly across borders. However, this digital transformation has also accelerated the spread of misinformation, creating challenges in discerning fact from falsehood. Misinformation—including fake news, propaganda, and unfounded rumors—has become a pervasive issue with serious implications for public health, politics, and social cohesion. This "infodemic" was particularly evident during events like the COVID-19 pandemic, where unverified claims proliferated, fueling confusion and exacerbating societal issues.

In response to the growing demand for fact-checking and misinformation countermeasures, Natural Language Processing (NLP) researchers have developed various automated systems to detect and verify online claims. These systems tackle misinformation through sub-tasks such as claim detection, claim verification, and check-worthiness assessment. Among these, Claim Span Identification (CSI) has emerged as an essential component of the fact-checking pipeline. CSI involves identifying specific segments within text that constitute claims, enabling downstream tasks like verification and validation.

While much progress has been made in claim span identification for English-language texts, multilingual environments present unique challenges. Different languages embody distinct structures, expressions, and cultural contexts, which can hinder the accuracy of claim span identification models trained on English data alone. Hindi, for instance, with its regional and political nuances, introduces complexities in accurately distinguishing claim-worthy segments in social media posts.

To address these issues, we focus on developing and testing a multilingual claim span identification framework that includes English and Hindi datasets. Our project leverages recent advances in NLP models, specifically transformer-based architectures, to improve the accuracy and robustness of claim span detection in both high-resource (English) and low-resource (Hindi) languages. By tailoring our approach to these languages, we aim to bridge the performance gap typically observed in multilingual tasks, ensuring that claim span identification remains effective across diverse linguistic settings.

The primary datasets used in this study is the ICPR-CSI dataset, each supporting multilingual analysis with claim span annotations. Both datasets feature real-world social media claims across a range of topics, with Hindi data primarily encompassing political discourse, while English data spans a broader array of themes.

2 Related Works

Related Work Claim span detection has emerged as a critical task in argument mining, facilitating advancements in fact-checking, misinformation detection, and discourse analysis. This section reviews significant contributions in this domain.

Early studies by Wang et al. (2020) laid the groundwork for neural sequence tagging methods to identify argumentative components. They utilized contextual embeddings and attention mechanisms, showing that task-specific representations significantly enhance performance in detecting claims and premises.

Research into multilingual argument mining has gained traction, particularly with the development of annotated datasets for low-resource languages. One such effort introduced an annotated corpus for Hindi, focusing on claims labeled using IO tagging. This initiative highlighted the need for creating culturally and linguistically diverse datasets to address the challenges of syntactic and semantic variability.

Transformer-based architectures, such as RoBERTa, have played a pivotal role in advancing claim detection methodologies. Studies leveraging pre-trained models fine-tuned with domain-specific data have achieved state-of-the-art performance. These models demonstrate robustness in detecting claims across various benchmarks and domains, particularly when paired with task-specific enhancements.

In morphologically rich languages like Hindi, models such as MuRIL have shown superior performance in claim span detection. The integration of MuRIL with sequence models like BiGRU and transformer encoders has proven effective, as demonstrated in work by Megha et al., achieving high precision and recall in argumentative tasks. Their approach highlights the importance of tailoring models to address language-specific challenges.

Furthermore, hybrid models combining multi-head attention mechanisms with transformer encoders have been successful in capturing long-range dependencies in argumentative texts. These models, as explored in recent research, significantly improve claim boundary identification.

The challenges of detecting claims in noisy and unstructured data, such as social media and online discourse, remain a significant focus. Fine-grained annotation schemas, particularly IO tagging, have been instrumental in addressing these challenges and refining the performance of various models.

Overall, advancements in transformer-based architectures, multilingual pre-trained models, and annotation methodologies have driven significant progress in claim span detection. However, challenges related to data scarcity, language variability, and domain-specific contexts persist, necessitating ongoing research and innovation.

3 Data and Methods

3.1 Dataset Description

The dataset used for this project is specifically designed for claim span detection tasks. It includes annotations marking the presence of claims in text, with precise start and end indices for claim spans. The dataset comprises the following columns:

- **Index:** A unique identifier for each instance in the dataset.
- **Claim-Index:** A numeric label representing distinct claims within the dataset.
- **Claim-Start and Claim-End:** Indices specifying the beginning and end of each claim span within the text.
- **Claim-Terms:** The actual text corresponding to the identified claim span.
- **Text-Tokens:** The tokenized version of the full text in which claims are identified.

The dataset contains diverse linguistic structures and contexts, reflecting the challenges of claim span detection in multilingual and noisy textual data.

3.1.1 Data Preprocessing

To prepare the dataset for model training, the following preprocessing steps were applied:

3.1.2 Tokenization:

- Utilized the SentencePiece tokenizer to split the text into subword units, ensuring compatibility with transformer models like MuRIL.
- Special care was taken to align the tokenized text with the annotated labels.

3.1.3 IO Labeling:

- Each token was labeled using the BIO scheme:
    - I: Inside the claim span.
    - O: Outside the claim span.
- This tagging scheme allows precise identification of claim boundaries.

3.1.4 Handling Special Cases:

- Addressed instances where tokens spanned across annotated claim boundaries by adjusting the labeling for subword tokens.
- Resolved ambiguities in overlapping or nested claims by assigning primary spans based on hierarchical priority.

3.2 Methodology

To effectively tackle the nuanced nature of this task, we integrated encoder models with sequence models, aiming to achieve a robust and high-dimensional representation of the input data. This hybrid approach leverages the strengths of both types of models to address the challenges inherent in token-level binary classification.

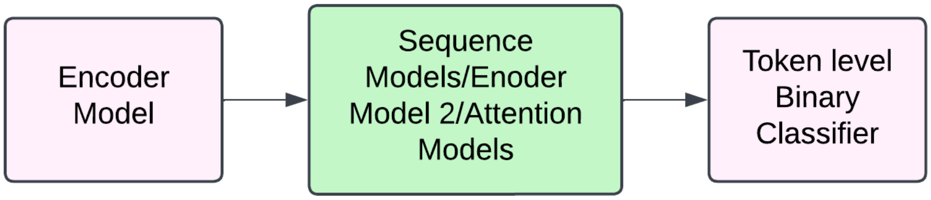
Encoder Model
Sequence Models/Enoder Model 2/Attention Models
Token level Binary Classifier

Figure 1. Base Architecture

- **Initial Ingestion:** The IO-tagged data underwent preprocessing to ensure consistency and readiness for model ingestion. These data were then fed into encoder models such as MuRIL or RoBERTa, which are known for their ability to extract contextual embeddings from raw text. The high-dimensional representations generated by these encoders served as the foundation for downstream tasks, effectively capturing global context and semantic nuances.

- **Sequence Models:** Following the encoder layer, we employed sequence models such as GRUs. These models are particularly effective at retaining temporal and sequential information, making them ideal for capturing localized, dense representations of claim spans. The sequence models enabled us to perform refined local transformations, ensuring that subtle patterns within the data were not overlooked. By combining the strengths of contextual embeddings with the sequential processing capabilities of these models, we enhanced the system's ability to differentiate intricate token-level features.

- **Attention Mechanisms:** To further improve the model's understanding of token relationships, we integrated attention mechanisms. These mechanisms allowed the model to focus on critical parts of the input sequence, dynamically assigning importance weights to different tokens. This was especially useful in tasks requiring nuanced analysis, as it ensured that relevant spans received greater emphasis.

- **Binary Classifier:** Since the task ultimately involved token-level binary classification, we designed a dedicated classification layer optimized for this purpose. This layer was tasked with predicting whether each token belonged to a specific class or not. We experimented with standard loss functions like binary cross-entropy, coupled with metric F1 score, to fine-tune and evaluate the classifier's performance.

- **Regularization and Optimization:** To mitigate overfitting, we incorporated regularization techniques such as dropout. Optimization was carried out using adaptive methodslike AdamW, ensuring efficient convergence during training.

- **Error Analysis and Iterative Refinement:** After initial training, we performed detailed error analysis to identify patterns in misclassified tokens. Insights from this analysis informed iterative model refinements, such as tweaking hyperparameters, augmenting training data, and enhancing preprocessing pipelines.

4 Results and Discussions

Results: The performance of the stacked models was evaluated using F1-Score on the test set. These are the results for the stacked architectures.

Model Architecture	F1-Score (English)	F1-Score (Hindi)
MuRIL + BiGRU	80.5	91
RoBERTa + TransformerEncoder	90.8	78.8
RoBERTa + Multi-Head Attention	64.7	54.5
MuRIL + TransformerEncoder	78.1	88.2
MuRIL + Multi-Head Attention	55.9	79.8

Table 1
Stacked Models Result

5 Conclusion

The experimental results demonstrate significant variability in model performance across both English and Hindi datasets, providing insights into the relative effectiveness of different model architectures. Several notable trends can be observed:

- **Language-Specific Trends:** The combination of MuRIL and BiGRU achieved the highest performance for Hindi (**91 F1 score**), highlighting the advantage of using MuRIL, a model pre-trained on Indian languages. Conversely, RoBERTa with Transformer Encoder excelled in English with an **F1 score of 90.8**, indicating that RoBERTa's generalized contextual embeddings are better suited for English text. This underscores the importance of selecting encoder models tailored to the linguistic characteristics of the target language.

- **Encoder Strengths:** Across all experiments, the choice of encoder significantly influenced performance. MuRIL consistently outperformed RoBERTa in Hindi, while RoBERTa demonstrated superior results for English, suggesting that language-specific pretraining can provide a competitive edge, particularly for non-English datasets.

- **Effectiveness of Sequence Models:** BiGRU, when paired with MuRIL, delivered robust performance in both languages, emphasizing the importance of capturing sequential and temporal dependencies. This combination proved particularly effective in Hindi, where the sequential patterns of claim spans are more nuanced.

- **Performance of Attention Mechanisms:** Attention mechanisms such as MHA (Multi-Head Attention) displayed mixed results. Architectures incorporating MHA (**RoBERTa+MHA** and **MuRIL+MHA**) performed comparatively lower than other combinations. This suggests that while attention mechanisms help capture token relationships, they may not be sufficient on their own and need to be complemented by strong contextual embeddings or sequential models.

- **General Observations on Transformer Encoders:** Architectures leveraging Transformer Encoder layers (**RoBERTa+TransformerEncoder** and **MuRIL+ TransformerEncoder**) demonstrated consistent and balanced performance across both languages. These results highlight the capability of Transformer-based architectures to generalize across different datasets, especially when paired with high-quality embeddings.

- **Task Complexity and Overfitting:** Models with suboptimal scores (**MuRIL+MHA** and **RoBERTa+MHA**) might have struggled with overfitting due to their inability to fully capture localized token dependencies. This reinforces the need for regularization techniques and iterative refinements during training.

Overall, the results reveal that an appropriate combination of pre-trained encoders, sequence models, and attention mechanisms can effectively tackle token-level binary classification tasks. Further research can explore fine-tuning hyperparameters, augmenting data, or employing ensemble methods to improve performance further.

.

References

Ahmed, K., **Khan, M. A.**, **Haq, I.**, **Al Mazroa, A.**, **Syam, M.**, **Innab, N.**, **Alajmi, M.**, and **Alkahtani, H. K.** (2024). *Social media's dark secrets: A propagation, lexical and psycholinguistic oriented deep learning approach for fake news proliferation.* Expert Systems with Applications, 255:124650.

Barrón-Cedeño, A., **Alam, F.**, **Chakraborty, T.**, **Elsayed, T.**, **Nakov, P.**, **Przybyła, P.**, **Struß, J. M.**, **Haouari, F.**, **Hasanain, M.**, **Ruggeri, F.**, and others (2024). *The clef-2024 checkthat! lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness.* In European Conference on Information Retrieval, pp. 449–458. Springer.

Bender, E. M., **Morgan, J. T.**, **Oxley, M.**, **Zachry, M.**, **Hutchinson, B.**, **Marin, A.**, **Zhang, B.**, and **Ostendorf, M.** (2011). Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 48–57, Portland, Oregon. Association for Computational Linguistics.

Bhatnagar, V., **Kanojia, D.**, and **Chebrolu, K.** (2022). Harnessing abstractive summarization for fact-checked claim detection. In Proceedings of the 29th International Conference on Computational Linguistics, pp. 2934–2945.

Chakrabarty, T., **Hidey, C.**, and **McKeown, K.** (2019). IMHO fine-tuning improves claim detection. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 558–563, Minneapolis, Minnesota. Association for Computational Linguistics.

Chhablani, G., **Sharma, A.**, **Pandey, H.**, **Bhartia, Y.**, and **Suthaharan, S.** (2021). NLRG at SemEval-2021 task 5: Toxic spans detection leveraging BERT-based token classification and span prediction techniques. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 233–242, Online. Association for Computational Linguistics.

Choi, E. C. and **Ferrara, E.** (2024). Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. In Companion Proceedings of the ACM on Web Conference 2024, pp. 1441–1449.

Coope, S., **Farghly, T.**, **Gerz, D.**, **Vulic, I.**, and **Henderson, M.** (2020). Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 107–121, Online. Association for Computational Linguistics.

Da San Martino, G., **Barrón-Cedeño, A.**, **Wachsmuth, H.**, **Petrov, R.**, and **Nakov, P.** (2020). SemEval-2020 task 11: Detection of propaganda techniques in news articles. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1377–1414, Barcelona (online). International Committee for Computational Linguistics.