Enhancing Textual Understanding: Automated Claim Span Identification in English, Hindi, Bengali, and CodeMix

Rudra Roy[⋆], **Pritam Pal**[†], **Dipankar Das**[†], **Saptarshi Ghosh**[‡] and **Biswajit Paul**[⋄]
[⋆]Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India
[†]Jadavpur University, Kolkata, India
[‡]Indian Institute of Technology, Kharagpur, India
[⋄]Centre for Artificial Intelligence and Robotics, DRDO, Bengaluru, India
{royrudra164, pritampal522, dipankar.dipnil2005, saptarshi.ghosh}@gmail.com, biswajit.cair@gov.in

Abstract

Claim span identification, a crucial task in Natural Language Processing (NLP), aims to extract specific claims from texts. Such claim spans can be further utilized in various critical NLP applications, such as claim verification, fact-checking, and opinion mining, among others. The present work proposes a multilingual claim span identification framework for handling social media data in English, Hindi, Bengali, and CodeMixed texts, leveraging the strengths and knowledge of transformer-based pre-trained models. Our proposed framework efficiently identifies the contextual relationships between words and precisely detects claim spans across all languages, achieving a high F1 score and Jaccard score. The dataset is available at: https://github.com/pritampal98/claim-span-multilingual

1 Introduction

The increasing adoption of low-cost and high-speed internet has led to the proliferation of social media platforms, such as Twitter and Facebook, with India alone containing over 462 million users as of January 2024 (Kemp, 2024). While these platforms facilitate open expression, they are also prone to the spread of various claims, some of which contain misinformation or fake content, posing significant societal risks.

A ‘claim’ is a minimal text segment (e.g., in tweets, news articles) expressing verifiable information that the author asserts as true. Due to the prevalence of harmful misinformation, it is essential to identify and check factual claims on social media effectively.

Artificial Intelligence (AI) and Natural Language Processing (NLP) have made it possible to perform tasks such as identifying texts that are eligible for claims, distinguishing between factual and opinionated statements, and verifying claims by extracting supporting evidence. However, taking a more detailed approach that identifies the specific phrases or sections within a sentence that represent a claim provides even greater value.

Our present research focuses on identifying these specific claim spans within sentences or texts. Motivated by the need to identify both factual information and misinformation in social media, our contributions are summarized as follows:

- We expanded the ICPR-CSI (Poddar et al., 2025) dataset, annotating 4,988 new texts with claim spans in English, Hindi, and the less-explored Bengali and Hindi-English CodeMixed texts.
- We develop a multilingual claim span identification framework using different transformer-based pre-trained models for the languages as mentioned above, utilizing two tagging schemes (Binary and B-I-O).

2 Related Work

Research in claim identification, spurred by advancements in deep learning and NLP, has evolved from simple claim detection (Rosenthal and McKeown, 2012; Chakrabarty et al., 2019; Pathak et al., 2020; Gupta et al., 2021; Wührl and Klinger, 2021; Sundriyal et al., 2021; Gangi Reddy et al., 2022a; Das et al., 2024) to check-worthiness assessment (Jaradat et al., 2018; Wright and Augenstein, 2020) to the more specific identifying factual claim spans (Sundriyal et al., 2022; Mittal et al., 2023).

Early work by Pathak et al. (2020) proposed a self-supervised, attention-based method for identifying claim-worthy sentences in fake news. More recently, frameworks for claim span identification in social media texts have emerged. Sundriyal et al. (2022) developed a transformer-based approach for English texts, while Mittal et al. (2023) proposed a multilingual framework covering five Indian languages and English. In 2023, the ‘CLAIMSCAN’

Proceedings of Recent Advances in Natural Language Processing, pages 1030–1035
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-118

shared task at FIRE 2023 (Sundriyal et al., 2023) further facilitated research with the release of a 7.5K claim span dataset. In 2024, the ICPR hosted a shared task focused on multilingual claim span identification (Poddar et al., 2025), which saw participation from over 40 teams. Among these teams, the transformer-based approach, specifically XLM-RoBERTa, achieved the highest results. In contrast, one team utilized an LLM-based approach, but their performance fell short of that of the transformer-based method.

Claim detection is also well-explored in news articles and specialized domains, particularly concerning COVID-19. Datasets like those proposed by Salek Faramarzi et al. (2023) include COVID-19-related tweets for check-worthiness tasks. Natural Language Inference (NLI) frameworks have been developed to identify contradictory claims about COVID-19 drug efficacy (Sosa et al., 2023), and synthetic datasets like EMCC aim to align medical claims with evidence (Hughes and Song, 2024). In news articles, methods for cross-lingual identification of political claims (Zaberer et al., 2023) and benchmarks like "NEWSCLAIMS" for attribute-aware claim detection have been introduced (Gangi Reddy et al., 2022b).

While claim identification has been widely studied across various domains, the specific task of identifying claim spans has received less attention. Existing research by Sundriyal et al. (2022) and Mittal et al. (2023) utilized sample sizes of approximately 7,500 and 7,000, respectively.

3 Dataset

The proposed claim span identification task primarily utilized the 'ICPR-2024 Multilingual Claim-Span Identification (ICPR24-CSI)' dataset (Poddar et al., 2025), which comprises ≈ 15K English and Hindi samples, each detailing text tokens and corresponding claim spans.

Further, we extended this dataset by manually annotating with 4,988 texts (English, Hindi, Bengali, and CodeMixed). These were collected from Twitter (focusing on Indian elections, farmer protests, and COVID-19 lockdown) and Bengali news headlines. Moreover, we integrated 1250 manually annotated texts with claim spans from the 'THAR (Targeted Hate Speech Against Religion)' Dataset (Sharma et al., 2024). The inclusion of THAR data was motivated by the presence of claims within hate speech, which often contribute to misinformation.

Annotation Strategy: Three undergraduate interns, proficient in English, Hindi, Bengali, and CodeMixed, performed the manual annotation of claims for the THAR subset and our in-house collected Twitter data. Annotators were guided by three core principles to tag claim spans in a text:

1. *Minimally meaningful*: Select the shortest text segment sufficient to express the claim.
2. *Verifiability*: The selected span must contain information whose truthfulness can be externally verified.
3. *Authorship*: For nested information (e.g., "X said that Y"), the claim pertains to what X said, not necessarily the veracity of Y.

Annotators received clear instructions and examples for each principle to ensure accurate claim span tagging.

Our final dataset, combining the ICPR-CSI data with our annotations, totals 1574 English, 1712 Hindi, 655 Bengali, and 1047 CodeMixed texts with claim spans. Examples from this dataset are shown in Figure 1, while its language distribution is detailed in Table 1.

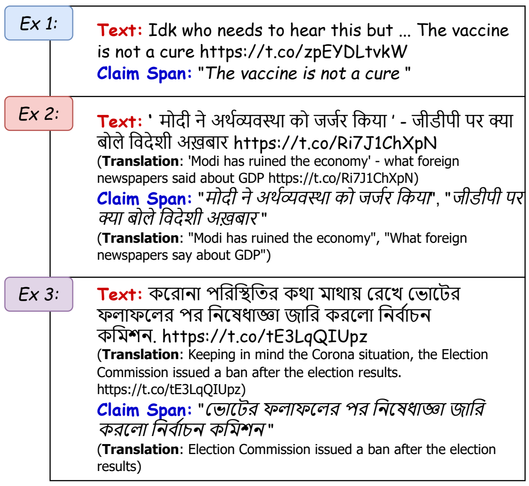

Figure 1: claim span examples from final dataset.

Language	Training	Validation	Testing
English	7573	500	1500
Hindi	7810	500	1500
Bengali	545	–	110
CodeMixed	971	–	100
Total	16,899	1000	3210

Table 1: Distribution of data in the final dataset

4 Methodology

We aim to develop a claim span identification framework, treating it as a word-level classification task. Given a sentence $S = [w_1, w_2, ..., w_n]$, the goal is to classify each word w_i as either a 'Claim' or 'Not-Claim' word. This can be represented as an output sequence $[c_1, c_2, ..., c_n]$, where $c_i \in \{0, 1\}$ for binary tagging (0 for 'Not-Claim', 1 for 'Claim'), or $c_i \in \{0, 1, 2\}$ for the B-I-O tagging scheme (0 for 'O', 1 for 'I', 2 for 'B'). The overall model flow diagram is depicted in Figure 2.

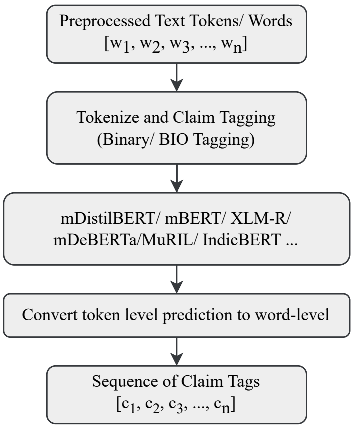

Figure 2: Flow diagram for claim span identification.

Text Preprocessing and Tokenization: Since most of the data were sourced from social media, some preprocessing steps were taken to reduce noise: (i) removal of escape sequences, (ii) URL standardization (replaced with 'http'), (iii) username anonymization (replaced with '@user'), and (iv) unknown tokens were replaced with `<unk>`.

Following preprocessing, texts were tokenized using the corresponding tokenizer for each pretrained model (e.g., BERT tokenizer for mBERT). Transformer tokenizers often split words into subtokens, which can obscure original word boundaries. To address this, we captured 'word_ids' at the time of the tokenization of texts. The 'word_ids' preserves the original positional index of each word, crucial for accurate claim word tagging.

Claim Tagging: Two schemes for assigning claim tags to words were used: binary tagging and B-I-O tagging. The binary tagging method, inspired by the ICPR24-CSI shared task, assigns tag '1' (Claim) to words within a claim span and tag '0' (No-Claim) to all others.

Along with binary tagging, we also adopted the B-I-O tagging scheme (Ramshaw and Marcus, 1995), commonly used in sequence labeling tasks like Named Entity Recognition. Here, 'B' (B-CLAIM) denotes the first word of a claim span, 'I' (I-CLAIM) marks subsequent words within the span, and 'O' (O-CLAIM) indicates words outside any claim span. This approach provides more granular boundary information than binary tagging.

Model Selection and Training: After claim tagging, all token IDs (numeric representation of tokens), attention masks, and tags were prepared. We selected a diverse set of transformer models, ranging from the lightweight mDistilBERT(Sanh et al., 2019) to the powerful mBERT(Devlin et al., 2018) and XLM-RoBERTa (base and large) (Conneau et al., 2019), and the highly effective mDeBERTa, which performs well in multilingual settings (He et al., 2021). Given our focus on Indian languages and CodeMixed texts, we also included MuRIL (base and large) (Khanuja et al., 2021), IndicBERT (Kakwani et al., 2020), and IndicBERTv2 (Doddapaneni et al., 2023). These Indic-language-specific models, trained on 12-24 Indian languages, are expected to capture more nuanced linguistic aspects in Hindi, Bengali, and Code-Mixed texts than general multilingual models.

The claim span identification framework was trained by providing token IDs and attention masks to the selected models as input, aiming for an accurate prediction of claim tags. We used the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 3e-5 and a weight decay of 0.01. All models were trained for three epochs with a batch size of 8.

Sub-word Token to Word-level: Following training, subword-level outputs from the transformer models were converted back to word-level, utilizing the 'word_ids' mapping provided by the tokenizer, effectively re-aligning subword predictions to their original words. This ensures that a single prediction is made for each original word, regardless of how it was tokenized into subwords.

5 Results

The claim span identification framework was evaluated using both Binary and B-I-O tagging schemes across English, Hindi, Bengali, and CodeMixed datasets. The performances were measured using macro F1-score and Jaccard similarity score [1], with detailed results shown in Table 2.

Binary Tagging: In English, the mDeBERTa model achieved the highest individual F1-score, which was 78.73. For Hindi, IndicBERTv2 demonstrated the best standalone performance with an F1-score of 83.45 and a Jaccard score of 71.77.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.jaccard_score.html

	Binary Tagging								B-I-O Tagging							
Model	English		Hindi		Bengali		CodeMix		English		Hindi		Bengali		CodeMix	
	F1.	Jacc.	F1.	Jacc.	F1.	Jacc.	F1.	Jacc.	F1.	Jacc.	F1.	Jacc.	F1.	Jacc.	F1.	Jacc.
ICPR-CSI	74.8	54.5	81.7	67.1	-	-	-	-	-	-	-	-	-	-	-	-
mDistilBERT	75.43	60.74	79.65	66.42	74.12	58.97	75.06	61.9	62.76	48.15	74.28	59.74	69.51	53.72	64.58	50.79
mBERT	75.87	61.32	80.64	67.73	72.11	56.6	73.36	59.81	64.72	50	75.4	61.15	68.65	52.82	64.38	50.25
XLM-R	78.4	64.67	83.28	71.51	74.4	59.25	77.68	64.85	66.28	51.66	78.1	64.64	71.06	55.58	64.51	50.64
XLM-R-Lg	78	64.06	82.78	70.82	73.69	58.35	74.23	60.78	**67.7**	**53.03**	79.03	65.81	70.21	54.25	68.28	53.89
mDeBERTa	**78.73**	**65.03**	83.37	71.68	73.55	58.17	**79.64**	**67.44**	67.57	52.9	79.32	66.2	69.1	52.95	67.28	52.66
MuRIL	77.13	62.87	82.76	70.78	74.21	58.99	75.54	62.15	65.93	51.09	78.45	65.01	69.4	53.32	67.2	52.93
MuRIL-Lg	77.74	63.72	83.39	71.68	**75.64**	**60.82**	74.99	61.7	66.96	52.32	78.91	65.65	**71.48**	**55.75**	67.31	52.88
IndicBERT	76.48	62.11	78.31	64.6	70.48	54.49	78.22	65.93	63.66	48.96	72.11	57.02	65.75	49.42	66.68	52.9
IndicBERTv2	78.06	64.15	**83.45**	**71.77**	74.74	59.69	77.51	64.74	66.38	51.73	79.05	65.91	69.65	53.63	**68.8**	**54.73**

Table 2: Results of claim span identification on different experimented models for binary and B-I-O tagging scheme.

In Bengali, the MuRIL-large model achieved the top individual F1-score of 75.64. When evaluating CodeMix data, mDeBERTa was also the best individual performer, with an F1-score of 79.64 and a Jaccard score of 67.44.

Our proposed frameworks consistently outperformed the state-of-the-art benchmarks set by the ICPR-CSI organizers for both English and Hindi. The best individual model, mDeBERTa, exceeded the English benchmark by 5.25%. IndicBERTv2 showed a 2.14% improvement over the Hindi benchmark.

B-I-O Tagging: Under the B-I-O tagging scheme, XLM-RoBERTa-large achieved the highest individual F1-score of 67.70 (Jaccard: 53.03) for English. For Hindi, the top-performing model was mDeBERTa, with an F1-score of 79.32 and a Jaccard score of 66.20.

In Bengali, the best individual model was MuRIL-large, which scored an F1 of 71.48 and a Jaccard score of 55.75. For CodeMix data, IndicBERTv2 was the leading model, achieving an F1-score of 68.70 and a Jaccard score of 54.73.

Observations: Upon experimenting with different pre-trained transformer models, a few key observations have been drawn:

First, while the best individual performances varied across languages and tagging schemes (e.g., mDeBERTa for English and CodeMix in Binary tagging, IndicBERTv2 for Hindi in Binary tagging and CodeMix in B-I-O tagging), overall model strength can be seen in their average performance. For the Binary tagging scheme, the mDeBERTa model yielded the best average F1-score of 78.82 (Jaccard: 65.58), followed by IndicBERTv2 (F1: 78.44, Jaccard: 65.09) and XLM-RoBERTa (F1: 78.44, Jaccard: 65.07). For B-I-O tagging, XLM-RoBERTa-large provided the best average performance (F1: 71.30, Jaccard: 56.74), closely followed by MuRIL-large (F1: 71.16, Jaccard: 56.65). These models therefore represent optimal choices for robust claim span identification across English, Hindi, Bengali, and CodeMix texts.

Second, regarding the B-I-O tagging scheme, larger variants of transformer models (XLM-RoBERTa-large, MuRIL-large) generally outperformed their base counterparts. Conversely, in the Binary tagging scheme, base variant transformer models tended to show better performance than their large variants.

6 Conclusion and Future Work

This paper presented a multilingual framework for claim span identification across English, Hindi, Bengali, and Code-Mixed texts, leveraging nine transformer-based pre-trained models. We investigated two tagging schemes: a binary scheme and a more granular B-I-O scheme.

Experimental results demonstrated that the mDeBERTa, IndicBERTv2, and XLM-RoBERTa showed strong average performance for binary tagging, while XLM-RoBERTa-large and MuRIL-large excelled under the B-I-O scheme. Notably, our frameworks surpassed existing state-of-the-art benchmarks for English and Hindi claim span identification.

For future work, we plan to expand our dataset further to strengthen our findings and observations, and incorporate additional low-resource languages to enhance the robustness and broader applicability of our claim span identification framework.

Acknowledgement

This work was supported by Defence Research and Development Organisation (DRDO), New Delhi, under the project "Claim Detection and Verification using Deep NLP: an Indian perspective".

References

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 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)*, pages 558–563, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Shankha Shubhra Das, Pritam Pal, and Dipankar Das. 2024. Unveiling the truth: A deep dive into claim identification methods. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 635–644, Tokyo, Japan. Tokyo University of Foreign Studies.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Yi R. Fung, Kevin Small, and Heng Ji. 2022a. A zero-shot claim detection framework using question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6927–6933, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi Fung, Kathryn Conger, Ahmed ELsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small, and Heng Ji. 2022b. NewsClaims: A new benchmark for claim detection from news with attribute knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6002–6018, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shreya Gupta, Parantak Singh, Megha Sundriyal, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. LESA: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3178–3188, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Anthony James Hughes and Xingyi Song. 2024. Identifying and aligning medical claims made on social media with medical evidence. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8580–8593, Torino, Italia. ELRA and ICCL.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, New Orleans, Louisiana. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Simon Kemp. 2024. Digital 2024: India — DataReportal – Global Digital Insights.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Shubham Mittal, Megha Sundriyal, and Preslav Nakov. 2023. Lost in translation, found in spans: Identifying claims in multilingual social media. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3887–3902, Singapore. Association for Computational Linguistics.

Archita Pathak, Mohammad Abuzar Shaikh, and Rohini Srihari. 2020. Self-supervised claim identification for automated fact checking. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 213–227, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Soham Poddar, Biswajit Paul, Moumita Basu, and Saptarshi Ghosh. 2025. Icpr 2024 competition on multilingual claim-span identification. In *Pattern Recognition. Competitions*, pages 134–144, Cham. Springer Nature Switzerland.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Sara Rosenthal and Kathleen McKeown. 2012. Detecting opinionated claims in online discussions. In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 30–37.

Noushin Salek Faramarzi, Fateme Hashemi Chaleshtori, Hossein Shirazi, Indrakshi Ray, and Ritwik Banerjee. 2023. Claim extraction and dynamic stance detection in covid-19 tweets. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 1059–1068, New York, NY, USA. Association for Computing Machinery.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Deepawali Sharma, Aakash Singh, and Vivek Kumar Singh. 2024. Thar- targeted hate speech against religion: A high-quality hindi-english code-mixed dataset with the application of deep learning models for automatic detection. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Daniel Sosa, Malavika Suresh, Christopher Potts, and Russ Altman. 2023. Detecting contradictory COVID-19 drug efficacy claims from biomedical literature. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–713, Toronto, Canada. Association for Computational Linguistics.

Megha Sundriyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Overview of the claimscan-2023: Uncovering truth in social media through claim detection and identification of claim spans.

Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Empowering the fact-checkers! automatic identification of claim spans on Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7701–7715, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Megha Sundriyal, Parantak Singh, Md. Shad Akhtar, Shubhashis Sengupta, and Tanmoy Chakraborty. 2021. Desyr: Definition and syntactic representation based claim detection on the web. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

Dustin Wright and Isabelle Augenstein. 2020. Claim check-worthiness detection as positive unlabelled learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 476–488, Online. Association for Computational Linguistics.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical Twitter posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 131–142, Online. Association for Computational Linguistics.

Urs Zaberer, Sebastian Pado, and Gabriella Lapesa. 2023. Political claim identification and categorization in a multilingual setting: First experiments. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 219–228, Ingolstadt, Germany. Association for Computational Linguistics.