Overview of the CLAIMSCAN-2023: Uncovering Truth in Social Media through Claim Detection and Identification of Claim Spans

Megha Sundriyal[1], Md Shad Akhtar[1] and Tanmoy Chakraborty[2]

[1]*IIIT Delhi, India*
[2]*IIT Delhi, India*

Abstract

A significant increase in content creation and information exchange has been made possible by the quick development of online social media platforms, which has been very advantageous. However, these platforms have also become a haven for those who disseminate false information, propaganda, and fake news. Claims are essential in forming our perceptions of the world, but sadly, they are frequently used to trick people by those who spread false information. To address this problem, social media giants employ content moderators to filter out fake news from the actual world. However, the sheer volume of information makes it difficult to identify fake news effectively. Therefore, it has become crucial to automatically identify social media posts that make such claims, check their veracity, and differentiate between credible and false claims. In response, we presented CLAIMSCAN in the 2023 Forum for Information Retrieval Evaluation (FIRE'2023). The primary objectives centered on two crucial tasks: Task A, determining whether a social media post constitutes a claim, and Task B, precisely identifying the words or phrases within the post that form the claim. Task A received 40 registrations, demonstrating a strong interest and engagement in this timely challenge. Meanwhile, Task B attracted participation from 28 teams, highlighting its significance in the digital era of misinformation.

Keywords

Claims, Social Media, Claim Detection, Claim Span Identification, Twitter, Misinformation, Fact-Checking

1. Introduction

The rapid growth of online social media platforms has facilitated a significant increase in content creation and information exchange, which has been highly beneficial. However, these platforms have also become a breeding ground for those who spread malicious rumors, fake news, propaganda, and misinformation. Claims play a vital role in shaping our understanding of the world, but unfortunately, they are often used by purveyors of fake news to deceive people. The COVID-19 *"Infodemic"* is a prime example of this phenomenon, which has resulted in the widespread dissemination of false information about politics and social issues, as well as fake medical claims [1]. To address this issue, social media giants hire content moderators to separate fake news from the real thing. However, the sheer volume of information makes it difficult to identify fake news effectively. As a result, automatically identifying posts on social media

Forum for Information Retrieval Evaluation, December 15-18, 2023, India

meghas@iiitd.ac.in (M. Sundriyal); shad.akhtar@iiitd.ac.in (M. S. Akhtar); chak.tanmoy.iit@gmail.com (T. Chakraborty)

Table 1
Task A: Representative examples of claims and non-claims.

Text	Claim
My heartfelt gratitude goes out to the men and women in uniform who did not back down from putting their lives in danger to save the lives of our citizens in difficult circumstances.	No
According to research into the dangers of cooking with aluminum foil, some of the toxic metal can contaminate food. This is especially true when cooking or heating spicy or acidic foods in foil. Aluminum levels in the body have been linked to osteoporosis and Alzheimer's disease.	Yes
Furthermore, health insurers should recognize alternative medicine as a treatment option because there is a chance of recovery.	No
Toothpaste Zaps Pimples. Don't pop your pimples! Daily Glow recommends applying toothpaste to a pimple before bed and washing it off with warm water when you wake up in the morning. Toothpaste draws impurities out of pores while also drying the skin and shrinking the pimple.	Yes

Table 2
Task B: Representative examples of claims and their claim spans.

Claim	Claim Span
According to research into the dangers of cooking with aluminum foil, some of the toxic metal can contaminate food. This is especially true when cooking or heating spicy or acidic foods in foil. Aluminum levels in the body have been linked to osteoporosis and Alzheimer's disease.	cooking with aluminum foil, some of the toxic metal can contaminate food.
Toothpaste Zaps Pimples. Don't pop your pimples! Daily Glow recommends applying toothpaste to a pimple before bed and washing it off with warm water when you wake up in the morning. Toothpaste draws impurities out of pores while also drying the skin and shrinking the pimple.	Toothpaste Zaps Pimples.

platforms containing such claims, verifying their validity, and distinguishing between credible and false claims has emerged as a critical research problem in NLP.

The concept of a claim, defined by Toulmin [2] as an assertion that deserves attention, is central to Argument Mining (AM). However, the segregation of claims is complex and challenging due to language structure and context variation across different sources. Differentiating between claims and non-claims is highly subjective and tricky, making it difficult for human annotators and advanced state-of-the-art neural models. Table 1 furnishes a few examples of claims and non-claims for more understanding. Although claim-detecting systems have advanced, there is still room for improvement in their precision and efficiency [3]. The dynamic nature of online social media platforms presents a significant challenge. New types of misinformation can emerge quickly, and keeping up with changing trends and patterns can take time. In addition to the challenges of efficiently identifying claims, another factor affecting the fact-checking task is extracting precise snippets of the claim from the entire social media post, which often contain extraneous irrelevant text [4]. Table 2 depicts claims and their corresponding claim spans.

Disentangling such argumentative units of misinformation from benign statements has numerous advantages, including performing downstream tasks like claim check-worthiness and verification, adding explainability to the coarse-grained claim detection task, and simplifying the fact-checking process for human fact-checkers. This task, however, is complex and requires overcoming technical obstacles such as language complexity and variability.

To this end, we present the CLAIMSCAN-2023, a shared task in the 2023 edition of the Forum for Information Retrieval Evaluation workshop. Through this shared task, we aim to develop systems that can effectively detect and identify claims within social media text. To accomplish this, we propose two sub-tasks:

- Task A Claim Detection: Given a social media post, the task is to identify whether or not the post contains a claim.
- Task B Claim Span Identification: Given a social media post containing a claim, the objective is to pinpoint the exact phrase of the post that constitutes the claim.

2. Background

The growth of online social media has greatly amplified the spread of misinformation, primarily through disseminating false claims. This presents a significant risk to online users, as misinformation can spread rapidly without any effective countermeasures in place. Consequently, tasks related to identifying and handling claims have gained considerable prominence within the field of Natural Language Processing (NLP), particularly as a crucial precursor to automated fact verification. Claims, as a core component of misinformation, have been the subject of extensive research from multiple perspectives in recent years. This includes areas such as Claim Detection [5, 6, 3], Claim Check-worthiness [7, 8], Claim Span Identification [4, 9], Claim Normalization [10], and Claim Verification [11, 12, 13, 14].

Pioneering efforts in the study of claims can be attributed to Bender et al. [15], who introduced the "Authority and Alignment in Wikipedia Discussions" corpus, which comprised around 365 discussions sourced from Wikipedia Talk Pages. This work garnered substantial attention from researchers focusing on claims and served as the cornerstone for the challenging field of automated claim detection. Over the last decade, the investigation of online claims has gained some traction within the NLP research community. A primary attempt was made by Rosenthal and McKeown [16]; they used a supervised approach based on sentiment and word-gram derivatives to mine claims from discussion platforms. Despite the fact that their work was limited to traditional machine-learning approaches, it laid the groundwork for future research in this field. Following research on claim detection, linguistically motivated features such as sentiment analysis, syntax, context-free grammar, and parse trees were heavily emphasized [17, 18, 19].

Given that the majority of studies at the time focused on domain-specific formal texts, Daxenberger et al. [20] addressed this limitation by conducting cross-domain claim detection across six diverse datasets, revealing both distinctive and shared features across different domains. Recent research has led to the use of Large Language Models (*LLMs*), which hold great promise. Chakrabarty et al. [5] demonstrated the power of fine-tuning with their ULMFiT language model, which was fine-tuned on a large Reddit corpus of approximately 5 million opinionated claims. A generalized claim detection model was proposed by Gupta et al. [6] that detects the presence of a claim in any online text, regardless of source. They worked with both structured and unstructured data by training a combination of linguistic encoders (part-of-speech and dependency trees) and a contextual encoder. Because language models incur significant computational overheads, Sundriyal et al. [3] addressed this issue and proposed a lighter framework that attempted to generate discernible feature spaces for individual classes while avoiding using LLMs and focusing on the definition-centric approach. Several computational social science researchers have expressed interest in the *CLEF*-2020 shared task organized by the *CheckThat! Lab* [21]. Williams et al. [22] won the task by fine-tuning the RoBERTa model [23], which was further strengthened by mean pooling and dropout. With their RoBERTa vectors supplemented with Twitter meta-data, Nikolov et al. [24] bagged second position.

The existing body of claim detection research primarily focuses on identifying claims at the sentence level rather than delving into the finer details of exact claim spans. As a result, a recent advancement in this field has moved away from broad, sentence-level claim identification models and toward more detailed, fine-grained claim span identification [4]. The idea of rationales was first presented by Zaidan et al. [25], who highlighted segments of the text that validated the conclusions of their label. They reported a significant improvement in performance after incorporating these rationales into the training process for sentiment classification of movie reviews. In the field of argumentation mining, Trautmann et al. [26] released the *AURC*-8 dataset, which includes token-level span annotations for the argumentative components of stance, as well as their corresponding label. The *SemEval* community has initiated coarse-grained span identification concerning other domains of argument mining such as toxic comments [27] and propaganda techniques [28]. These shared tasks amassed many solutions constituting transformers [29], convolutional neural networks [30], data augmentation techniques [31, 32, 33], and ensemble frameworks [34, 35]. Wührl and Klinger [36] compiled a corpus of around 1200 biomedical-related tweets with claim phrases. Apart from English, argument extraction has also been examined for other languages like Greek [37, 38] and German [39]. In a recent study conducted by Sundriyal et al. [4], a systematic approach was presented for identifying claim spans within social media posts. Additionally, they created an extensive Twitter corpus manually annotated specifically for this task.

3. Tasks Description and Settings

CLAIMSCAN-2023 shared task consists of two sub-tasks: Claim Detection and Claim Span Identification. Participants were free to engage in one or both sub-tasks.

Task A (Claim Detection): Given a social media post, the objective is to identify whether a claim is present within a provided post or not. This task can be quite demanding, as claims exhibit diverse structures and can be concealed within extensive text segments. Hence, the system needs to discern patterns and linguistic cues that are indicative of claims, which may encompass assertive language, explicit statements on a topic, and allusions to supporting evidence or sources.

Task B (Claim Span Identification): Following the initial determination of whether the post contains a claim, the subsequent step entails pinpointing the precise span of the claim within the post. It is crucial for the system to precisely identify the specific words or phrases that form the claim, as this information plays a pivotal role in assessing its accuracy during the fact-checking process.

4. Datasets

To accomplish Task A (Claim Detection), we utilize a publicly available large-scale claim detection dataset developed and curated for tweets [6]. The dataset was manually annotated extensively using carefully crafted guidelines, yielding a collection of $9,894$ tweets labeled as either containing a claim or not containing a claim. The statistics of the dataset are detailed in Table 3. For Task B (Claim Span Identification), we use the CURT dataset, which contains $9,458$ claim spans from $7,555$ tweets [4]. Table 4 contains the dataset statistics and details. This dataset has also been annotated manually, with each span identified and tagged using the *BIO* (Begin-Inside-Outside) encoding scheme [40], as shown in Table 5. This tagging scheme indicates whether each word in the tweet is within a claim span and, if so, whether at the start or end of the span.

Table 3
Task A: Statistics of claim detection dataset.

Dataset	Claim	Non-claim
Train set	7354	1055
Test set	1296	189
Overall	8650	1244

Table 4
Task B: Statistics of claim span identification dataset.

Dataset	Train	Test	Validation
Total no. of claims	6044	755	756
Avg. length of tweets	27.40	26.93	27.29
Avg. length of spans	10.90	10.97	10.71
No. of span per tweet	1.25	1.20	1.27
No. of single span tweets	4817	629	593
No. of multiple span tweets	1201	121	161

We took great care in developing annotation guidelines for both tasks, which went through several iterations and have already been published in two highly regarded peer-reviewed conferences. In addition, to ensure the quality of the data, we conducted pilot studies and enlisted human annotators with a strong understanding of claims and who are active social

Table 5
A few examples of social media posts from CURT dataset [4] and their corresponding *BIO* tags depicting claim spans.

Text	**Span**
@mcford77 @floradoragirl Exactly. that is the point. Home Schooling prevents loads of #Coronavirus deaths.	*{O, O, O, O, O, O, O, B, I, I, I, I, I, I}*
@JoeySalads Zero. #Covid19 is a hoax. The dead people died of something else. Where are the rest of the Corpses? If #coronavirus is real, then NYC would not be the greatest hit spot of DEATH from it in the world by a factor of five. What about Mexico City? Sydney?	*{O, O, B, I, I, I, B, I, I, I, I, I, I, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O}*

media users to manually annotate the datasets. This rigorous process helps to ensure data accuracy and reliability, resulting in more robust and reliable models. More details about the datasets can be found in Gupta et al. [6] and Sundriyal et al. [4].

5. Evaluation Metrics

The evaluation metric for both tasks is the F1 score. For Task A, we compute **Macro-F1** scores using *Scikit-learn* Library in Python used by the existing systems for claim detection [6, 3, 5]. For Task B, as the final labels for spans follow the *BIO* tagging notation, our task becomes a sequence labeling task. We compute **Token-F1** scores following existing span detection methods [27, 4]. Each team was allowed a maximum of 10 submissions, and the best scores obtained on test data were used for the leaderboard.

6. Participating Systems and Results

Task A received 40 registrations, and Task B received 28 registrations. Out of these 6 teams submitted their official runs for Task A, while 4 submitted for Task B. We first describe the teams that submitted system description papers.

- **Team NLytics[41]:** Team NLytics participated in both subtasks. For Task A, they fine-tuned the RoBERTa model [23] using RoBERTaForSequenceClassification, optimizing it with a regression loss (Binary Cross-Entropy Loss). They employed the AdamW optimizer with an initial learning rate of 2e-5. The optimizer followed a schedule where the learning rate increased linearly from 0 to the initial rate during a warm-up period and then decreased linearly to 0. The training process encompassed 20 training epochs. In Task B, they utilized RoBERTa and added a layer of linear-chain Conditional Random Field (CRF) [42]. As RoBERTa operates with byte pair encoding (BPE) units, while CRF requires whole words, only the initial tokens of words were used as input to the CRF, with any word continuation tokens being excluded. The training was started with 20 epochs, with an early stopping callback monitoring the model's performance on the validation set.

- **Team mjs227[43]:** Team mjs277 participated only in Task B. For identifying claim spans, they used the positional transformer architecture. The positional transformer is a transformer encoder architecture variant that uses a position-sensitive attention mechanism called positional attention. The underlying language model in their proposed model was RoBERTa$_{BASE}$ [23].
- **Team CODE[41]:** Team CODE participated in both subtasks. In Task A, they fine-tuned a BERT-based model [44] optimized for sequence classification and trained for 5 epochs. They utilized a binary cross-entropy loss (BCE loss) and employed an Adam optimizer for this task. In Task B, they employed the RoBERTa model and conducted fine-tuning to predict a binary label (0 or 1) for every token, indicating whether the token is associated with a claim or not. Instead of using the IOB tag set, they adhered to IO tags. Their model underwent training for a duration of 4 epochs, and to eliminate noise, they excluded instances with claim spans consisting of fewer than three words.

Table 6
Task A results for the best run per team based on macro-F1 scores.

Rank	Name	Macro-F1
1	NLytics	0.7002
2	bhoomeendra	0.6900
3	amr8ta	0.6678
4	CODE	0.6526
5	michaelibrahim	0.6324
6	pakapro	0.4321

The official results for Task A are presented in Table 6. Among the six participating teams, Team NLytics clinched the top position, attaining a noteworthy macro-F1 score of 0.7002. Following closely, Team bhoomeendra secured the second position.[1] Second position was bagged by Team amr8ta.[1] In the fourth spot for Task A was Team CODE, achieving a macro-F1 score of 0.6526. The fifth and sixth positions were occupied by Team michaelibrahim and Team pakapro, with macro-F1 scores of 0.6324 and 0.4321, respectively.[1] It's worth noting the substantial margin between the top-performing team and the rest.

The official results for Task B are in Table 7. Team mjs277 achieved the highest ranking among all participating teams, with a token-F1 of 0.8344. To identify claim spans, they harnessed the positional transformer architecture, resulting in a substantial enhancement of their model's performance. Team bhoomeendra secured the second position in the task, achieving a token-F1 score of 0.8030[1]. Team NLytics attained the third spot by fine-tuning a RoBERTa model for predicting BIO tags for each token in the input sentence, complementing it with a Conditional Random Field (CRF) layer. In fourth place was Team CODE, who opted for *IO* tags instead of *BIO* tags to signify whether a token was part of the claim or not.

[1] They did not release their system description papers.

Table 7
Task B results for the best run per team based on token-F1 scores.

Rank	Name	Token-F1
1	mjs227	0.8344
2	bhoomeendra	0.8030
3	NLytics	0.7821
4	CODE	0.5714

7. Conclusion

We presented the first edition of the CLAIMSCAN-2023 shared task. This shared task encompassed two vital subtasks within the fact-checking process, ranging from detecting claims in social media posts to determining the exact claim spans. These tasks collectively contribute to developing technology that aids human fact-checkers in their endeavors. We witnessed significant participation, with Task A drawing 40 registrations and Task B garnering 28 registrations. A total of 6 teams and 4 teams submitted official runs for Tasks A and B, respectively. We discussed the tasks and main findings of the three participating teams who submitted their systems based on their system description papers. We look forward to enriching our datasets with more examples, diverse information sources, and languages. Our overarching objective is to share our insights and inspire researchers to bridge the gaps in the field, ultimately enhancing the effectiveness of fact-checking systems and contributing to a safer online environment. In the future, we also aim to expand the scope of our task to encompass a broader range of modalities, such as images.

References

[1] S. B. Naeem, R. Bhatti, The covid-19 'infodemic': a new front for information professionals, Health information and libraries journal 37 (2020) 233—239. URL: https://europepmc.org/articles/PMC7323420. doi:10.1111/hir.12311.

[2] S. E. Toulmin, The uses of argument, Cambridge university press, 2003.

[3] M. Sundriyal, P. Singh, M. S. Akhtar, S. Sengupta, T. Chakraborty, Desyr: definition and syntactic representation based claim detection on the web, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 1764–1773.

[4] M. Sundriyal, A. Kulkarni, V. Pulastya, M. S. Akhtar, T. Chakraborty, Empowering the fact-checkers! automatic identification of claim spans on Twitter, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 7701–7715. URL: https://aclanthology.org/2022.emnlp-main.525.

[5] T. Chakrabarty, C. Hidey, K. McKeown, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 558–563. URL: https://aclanthology.org/N19-1054. doi:10.18653/v1/N19-1054.

[6] S. Gupta, P. Singh, M. Sundriyal, M. S. Akhtar, T. Chakraborty, 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, Association for Computational Linguistics, Online, 2021, pp. 3178–3188. URL: https://www.aclweb.org/anthology/2021.eacl-main.277.

[7] I. Jaradat, P. Gencheva, A. Barrón-Cedeño, L. Màrquez, P. Nakov, 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, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 26–30. URL: https://aclanthology.org/N18-5006. doi:10.18653/v1/N18-5006.

[8] D. Wright, I. Augenstein, Claim check-worthiness detection as positive unlabelled learning, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 476–488. URL: https://aclanthology.org/2020.findings-emnlp.43. doi:10.18653/v1/2020.findings-emnlp.43.

[9] S. Mittal, M. Sundriyal, P. Nakov, Lost in translation, found in spans: Identifying claims in multilingual social media, arXiv:2310.18205 (2023).

[10] M. Sundriyal, T. Chakraborty, P. Nakov, From chaos to clarity: Claim normalization to empower fact-checking, arXiv:2310.14338 (2023).

[11] S. Zhi, Y. Sun, J. Liu, C. Zhang, J. Han, Claimverif: A real-time claim verification system using the web and fact databases, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 2555–2558. URL: https://doi.org/10.1145/3132847.3133182. doi:10.1145/3132847.3133182.

[12] A. Hanselowski, H. Zhang, Z. Li, D. Sorokin, B. Schiller, C. Schulz, I. Gurevych, UKP-athene: Multi-sentence textual entailment for claim verification, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 103–108. URL: https://aclanthology.org/W18-5516. doi:10.18653/v1/W18-5516.

[13] A. Soleimani, C. Monz, M. Worring, Bert for evidence retrieval and claim verification, Advances in Information Retrieval 12036 (2020) 359.

[14] M. Sundriyal, G. Malhotra, M. S. Akhtar, S. Sengupta, A. Fano, T. Chakraborty, Document retrieval and claim verification to mitigate covid-19 misinformation, in: Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations, 2022, pp. 66–74.

[15] E. M. Bender, J. T. Morgan, M. Oxley, M. Zachry, B. Hutchinson, A. Marin, B. Zhang, M. Ostendorf, Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages, in: Proceedings of the Workshop on Language in Social Media (LSM 2011), Association for Computational Linguistics, Portland, Oregon, 2011, pp. 48–57. URL: https://www.aclweb.org/anthology/W11-0707.

[16] S. Rosenthal, K. McKeown, Detecting opinionated claims in online discussions, in: Proceedings of the 2012 IEEE Sixth International Conference on Semantic Computing, ICSC '12, IEEE Computer Society, USA, 2012, p. 30–37. URL: https://doi.org/10.1109/ICSC.2012.59.

doi:10.1109/ICSC.2012.59.

[17] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, N. Slonim, Context dependent claim detection, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 1489–1500.

[18] M. Lippi, P. Torroni, Context-independent claim detection for argument mining, in: Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015, pp. 185–191.

[19] R. Levy, S. Gretz, B. Sznajder, S. Hummel, R. Aharonov, N. Slonim, Unsupervised corpus–wide claim detection, in: Proceedings of the 4th Workshop on Argument Mining, 2017, pp. 79–84.

[20] J. Daxenberger, S. Eger, I. Habernal, C. Stab, I. Gurevych, What is the essence of a claim? cross-domain claim identification, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2055–2066. URL: https://aclanthology.org/D17-1218. doi:10.18653/v1/D17-1218.

[21] A. Barrón-Cedeno, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media, in: European Conference on Information Retrieval, Springer, Nature Publishing Group, 2020, pp. 499–507.

[22] E. Williams, P. Rodrigues, V. Novak, Accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models, arXiv:2009.02431 (2020).

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv:1907.11692 (2019).

[24] A. Nikolov, G. D. S. Martino, I. Koychev, P. Nakov, Team alex at clef checkthat! 2020: Identifying check-worthy tweets with transformer models, arXiv:2009.02931 (2020). arXiv:2009.02931.

[25] O. Zaidan, J. Eisner, C. Piatko, Using "annotator rationales" to improve machine learning for text categorization, in: Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference, 2007, pp. 260–267.

[26] D. Trautmann, J. Daxenberger, C. Stab, H. Schütze, I. Gurevych, Fine-grained argument unit recognition and classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 9048–9056. doi:https://doi.org/10.1609/aaai.v34i05.6438.

[27] J. Pavlopoulos, J. Sorensen, L. Laugier, I. Androutsopoulos, SemEval-2021 task 5: Toxic spans detection, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 59–69. URL: https://aclanthology.org/2021.semeval-1.6. doi:10.18653/v1/2021.semeval-1.6.

[28] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 task 11: Detection of propaganda techniques in news articles, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1377–1414. URL: https://aclanthology.org/2020.semeval-1.186.

[29] G. Chhablani, A. Sharma, H. Pandey, Y. Bhartia, S. Suthaharan, 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), Association for Computational Linguistics, Online, 2021, pp. 233–242. URL: https://aclanthology.org/2021.semeval-1.27. doi:10.18653/v1/2021.semeval-1.27.

[30] S. Coope, T. Farghly, D. Gerz, I. Vulić, M. Henderson, 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, Association for Computational Linguistics, Online, 2020, pp. 107–121. URL: https://aclanthology.org/2020.acl-main.11. doi:10.18653/v1/2020.acl-main.11.

[31] J. Rusert, Nlp_uiowa at semeval-2021 task 5: Transferring toxic sets to tag toxic spans, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 881–887.

[32] R. Palliser-Sans, A. Rial-Farràs, Hle-upc at semeval-2021 task 5: Multi-depth distilbert for toxic spans detection, arXiv:2104.00639 (2021).

[33] K. Pluciński, H. Klimczak, Ghost at semeval-2021 task 5: Is explanation all you need?, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 852–859.

[34] Q. Zhu, Z. Lin, Y. Zhang, J. Sun, X. Li, Q. Lin, Y. Dang, R. Xu, HITSZ-HLT at SemEval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, 2021.

[35] V. A. Nguyen, T. M. Nguyen, H. Q. Dao, Q. H. Pham, S-nlp at semeval-2021 task 5: An analysis of dual networks for sequence tagging, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 888–897.

[36] A. Wührl, R. Klinger, Claim detection in biomedical Twitter posts, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 131–142. URL: https://aclanthology.org/2021.bionlp-1.15. doi:10.18653/v1/2021.bionlp-1.15.

[37] T. Goudas, C. Louizos, G. Petasis, V. Karkaletsis, Argument extraction from news, blogs, and social media, in: Hellenic Conference on Artificial Intelligence, Springer, 2014, pp. 287–299.

[38] C. Sardianos, I. M. Katakis, G. Petasis, V. Karkaletsis, Argument extraction from news, in: Proceedings of the 2nd Workshop on Argumentation Mining, 2015, pp. 56–66.

[39] I. Habernal, I. Gurevych, Argumentation mining in user-generated web discourse, Computational Linguistics 43 (2017) 125–179.

[40] L. Ramshaw, M. Marcus, Text chunking using transformation-based learning, in: Third Workshop on Very Large Corpora, 1995. URL: https://aclanthology.org/W95-0107.

[41] A. Pritzkau1, J. Waldmüller, O. Blanc, M. Geierhos, U. Schade, Current language models' poor performance on pragmatic aspects of natural language, in: Proceedings of the CEUR Workshop Proceedings, Goa, India, CEUR, 2023.

[42] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers

Inc., San Francisco, CA, USA, 2001, p. 282–289. URL: https://openreview.net/forum?id=HkbzGjZOZB.

[43] M. Sullivan, N. Madani, S. Saha, R. Srihari, Positional transformers for claim span identification, in: Proceedings of the CEUR Workshop Proceedings, Goa, India, CEUR, 2023.

[44] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805 (2018).