JU_NLP@M&S at CheckThat! 2025: Automated Claim Extraction and Normalization for Misinformation Detection in Social Media Content

Notebook for the CheckThat! Lab at CLEF 2025

Malobika Mondal[1,†], Soumodeep Saha[1,*,†], Dipanjan Saha[1] and Dipankar Das[1,†]

[1]Dept of Computer Science and Engineering, Jadavpur University, Kolkata, West BengaL, India

Abstract

The rapid spread of misinformation on social media has intensified the need for automated tools that can identify and reformulate check-worthy claims in a clear and verifiable manner. This paper presents our approach to Task 2: Claims Extraction Normalization of the CLEF 2025 CheckThat! Lab, which focuses on the extraction and normalization of factual claims from noisy, user-generated content. We frame the problem as a monolingual sequence-to-sequence generation task and deploy a fine-tuned BART-Large transformer model to perform claim normalization.

The model is trained on the CLAN dataset comprising 6,388 annotated post-claim pairs, where each post is associated with one or more expert-generated normalized claims. Our preprocessing strategy includes careful tokenization, length normalization, and special handling of decoder inputs to facilitate accurate loss computation. Training is conducted using Hugging Face's Seq2SeqTrainer with mixed-precision optimization, beam search decoding, and ROUGE-based evaluation metrics. We compare the performance of BART-Large against baseline models including T5-Small and Pegasus. The fine-tuned BART-Large model achieves the best performance with a METEOR of 0.3098, significantly outperforming other models. Error analysis reveals challenges in negation handling, sarcasm detection, and multilingual noise, suggesting future avenues for enhancement. Our findings demonstrate that fine-tuned, transformer-based architectures are highly effective for claim normalization and hold promise for scalable, language-agnostic fact-checking systems designed combat misinformation online.

Keywords

Claim Extraction, Claim Normalization, BART-Large, Social Media, Fact-Checking, Misinformation, METEOR,

1. Introduction

In the age of social media and online information sharing, the rapid spread of misinformation seriously threatens public understanding and trust. Platforms like Twitter (now X), Facebook (now Meta), and Instagram have become hotspots for user-generated content, where opinions, news, and rumors intermingle freely. Among these, identifying factual claims—statements that can be verified as true or false—is crucial for the work of journalists, researchers, and automated fact-checking systems. However, the informal, unstructured, and often ambiguous nature of social media content makes this task significantly challenging. Manual identification and verification of claims in such noisy environments are both time-consuming and resource-intensive, creating the need for automated tools that can assist in this process.

Claim extraction and normalization refer to two interconnected tasks aimed at making misinformation detection more efficient:

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

malobika22ju@gmail.com (M. Mondal); soumodeepsahaa@gmail.com (S. Saha); sahadipanjan6@gmail.com (D. Saha); dipankar.dipnil2005@gmail.com (D. Das)

https://github.com/malobikacoder98 (M. Mondal); https://github.com/SoumodeepSaha (S. Saha); https://cse.jadavpuruniversity.in/faculty/dipankar-das (D. Das)

0009-0004-7501-501X (M. Mondal); 0009-0000-9387-6631 (S. Saha)

- **Claim Extraction:** This task involves identifying specific assertions or statements made within a social media post that can be verified. These claims often appear in the form of rumors, factual statements, or opinions that need validation. For example, in a post claiming, "The US government is investigating the war crimes in Afghanistan", the claim is that "The US government is investigating war crimes".
- **Claim Normalization:** Once claims are extracted, the next challenge is to simplify and rephrase them into a more clear, concise, and unambiguous form. Normalization ensures that claims are interpretable and ready for verification. This is particularly important for social media posts that may contain slang, sarcasm, or informal language that can obscure the true meaning of the claim. For example, the informal claim "The US might be looking into war crimes, who knows?" would be normalized into "The US is investigating war crimes in Afghanistan."

In this task, the goal is to transform noisy, complex, and often ambiguous social media posts into clear, well-defined claims that are easier to validate. This normalization process is critical for creating claims that automated systems or human fact-checkers can evaluate for accuracy and truthfulness.

To address these challenges, the **CLEF 2025 CheckThat! Lab** introduces *Task 2: Claim Extraction and Normalization*[1] [1], which focuses on the development of systems that can automatically identify and rephrase check-worthy claims from social media posts. This involves two key subtasks: first, extracting one or more factual claims from a post, and second, normalizing them—rephrasing the claims in a clear, concise, and unambiguous manner suitable for verification. For example, a tweet filled with slang, sarcasm, or mixed-language expressions might need to be reformulated into a standardized sentence that expresses the core assertion clearly. This normalization step is critical in ensuring that claims are interpretable and can be efficiently processed by automated fact-checking pipelines or human annotators.

A distinguishing feature of this task is its multilingual and inclusive approach. It spans 20 languages, including English, Arabic, Bengali, Hindi, and Tamil, reflecting the global nature of the misinformation problem. While the task encourages language-agnostic or language-specific models for claim extraction and normalization, this paper focuses solely on the English language track. The approach described here contributes to the fight against misinformation in English-speaking communities, helping to advance research in natural language understanding and promoting more scalable and effective fact-checking systems[2] in the process. With the availability of annotated datasets and a standardized evaluation framework, this task offers a robust platform for advancing claim-centric technologies.

2. Related Work

The task of claim normalization has been explored in various ways, focusing on different aspects such as claim detection[3], claim check-worthiness estimation[4], and claim extraction[5]. Previous research has predominantly worked on identifying claims and their verifiability, which is closely tied to fact-checking processes. Notable early work on claim detection is included in this article[6], which curated the AAWD corpus for claim detection, and later studies in the article[7], which expanded the domain to include claim identification across different topics.

Recent approaches in this area have incorporated large language models (LLMs), which have shown promise in improving claim detection and extraction in the article [8] and [9]. These models use linguistically motivated features like sentiment and syntax, which are essential for extracting claims from structured or semi-structured texts. However, these methods often fall short when handling the complex and noisy nature of social media data, which often requires abstractive claim extraction.[10]

Text summarization techniques[11], which condense lengthy documents into shorter summaries, have shown potential for solving problems related to claim normalization. It has been observed that in article[12] and [13] related to faithfulness in summarization have been taken care of. However, these methods do not focus on ensuring verifiability or factual consistency, which are critical for fact-checking.

[1] https://checkthat.gitlab.io/clef2025/task2/

On the other hand, claim normalization requires a more focused approach where the generated claims are not just concise but also self-contained and easily verifiable.

In line with text summarization, controlled summarization methods as given in the article [14] provide the ability to fine-tune summary attributes such as length and abstraction. These methods, however, still face challenges in producing summaries that retain factual accuracy. In contrast, claim normalization prioritizes verifiability, which sets it apart from general-purpose summarization tasks.

Our work extends the claim detection[15] and claim extraction fields by introducing a new challenge: claim normalization, which goes beyond summarization. Unlike previous models that focused on text condensation, our task aims to simplify complex and noisy social media posts into concise claims that fact-checkers can directly verify. We propose a novel approach, Check-worthiness Aware Claim Normalization (CACN), which integrates chain-of-thought reasoning and claim check-worthiness estimation to improve claim extraction from unstructured social media posts. This approach effectively adapts large language models to the specific needs of claim normalization. Moreover, we introduce the task of claim normalization for the verification of political claims, as detailed in the article.

3. Task Description

The task involves processing noisy and unstructured social media posts to extract specific and verifiable claims. These claims, often embedded in ambiguous or misleading content, need to be normalized—simplified into a more precise and easily understandable form. The objective is to bridge the gap between raw, informal social media data and structured, factual statements that can be efficiently verified by automated systems or manual fact-checkers. This process of transforming unstructured content into a clear, normalized claim enhances the accuracy and speed of the verification process, contributing to the mitigation of misinformation on social media platforms.

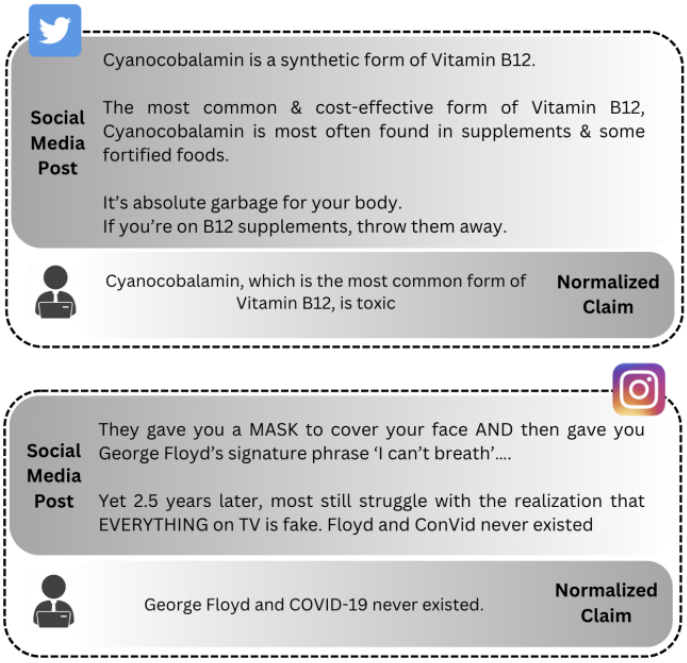

Figure 1: Illustration of the Claim Normalization task, highlighting the normalized claims authored by fact-checkers for social media posts from distinct social media platforms.

4. Dataset Description

To support the task of claim normalization, we utilize the **CLAN (Claim Normalization)** dataset, introduced by Sundriyal et al. (2023). This dataset [2] comprises **6,388 instances** of real-world social media posts, each paired with one or more **normalized claims**. These normalized claims are simplified, fact-checkable versions of the original posts, curated by professional fact-checkers as part of the verification process.

The dataset addresses the limitations of traditional summarization corpora by focusing specifically on extracting the **central verifiable assertion** from noisy, unstructured social media content. This emphasis enables downstream fact-checking systems to operate with greater efficiency and precision.

4.1. Data Sources

The CLAN dataset was constructed by collecting fact-checked claims and associated posts from two primary sources: *Google Fact-Check Explorer API* and *ClaimReview Schema*. Only the English-language posts were retained. All the non-textual entries, such as images or videos, were excluded to maintain textual uniformity and ensure relevance for language-based models.

4.2. Dataset Statistics

The dataset is split into training, validation, and test sets, with the following statistics:

Table 1
Statistics of the dataset used.

Split	Instances	Avg. Post Length (words)	Avg. Claim Length (words)
Train	5,341	39.52	16.47
Validation	594	37.12	17.24
Test	453	57.97	15.41
TOTAL	**6,388**	**44.87**	**16.37**

Notably, the test set includes multiple reference normalized claims per post to capture variability in how different annotators might distill the central assertion.

4.3. Data Characteristics

The dataset exhibits low cosine similarity between posts and normalized claims, confirming the claim normalization goes beyond extractive summarization. Rather than merely reducing the post length, the task involves abstractively transforming verbose, sometimes misleading content into clear, concise, and verifiable claims.

Representative examples are included in the original work, such as:

- **Post:** *Cyanocobalamin is a synthetic form of Vitamin B12...If you're on B12 supplements, throw them away.*
- **Normalized Claim:** *Cyanocobalamin, the most common form of Vitamin B12, is toxic.*

Such instances demonstrate the dataset's focus on real-world misinformation, public health concerns, and viral narratives on platforms like Twitter and Facebook.

[2] https://gitlab.com/checkthat_lab/clef2025-checkthat-lab/-/tree/main/task2/data

5. Methodology

The proposed methodology for the Claim Extraction and Normalization shared task is organized into four key modules: *data pre-processing*, *tokenization*, *model training*, and *decoding*. These modules are built around the fine-tuned **BART-large**[16] transformer model[3]. BART (Bidirectional and Auto-Regressive Transformers) is a state-of-the-art model for sequence-to-sequence tasks, pre-trained using a denoising autoencoder objective. It combines BERT's [17] bidirectional encoder, which is excellent for understanding context from both directions of a sequence, with GPT's[18] left-to-right decoder, which is efficient for generating coherent outputs. This unique combination makes BART highly suitable for tasks involving text generation and transformation, where both understanding and generating fluent text are critical.

The pre-training on large amounts of data allows the model to learn a wide range of linguistic patterns, which is why it excels in tasks like claim extraction and normalization. Specifically, we fine-tune BART-large on the task-specific dataset, which helps the model adapt to the particularities of claim normalization from noisy and unstructured social media content.

We approach this problem as a monolingual text-to-text generation task, where noisy or unstructured claims are transformed into their canonical, normalized forms. This task involves not just extracting a claim but also rephrasing it into a clearer, more verifiable statement, a challenge that requires both understanding the original message and generating a faithful, readable version.

5.1. Data Preprocessing and Tokenization

The first module handles data preprocessing and tokenization. Each input claim and its corresponding normalized output are tokenized using the BART tokenizer, which implements byte-level byte pair encoding (BPE)[19]. To prepare the text for the model, we ensure that each input has a uniform length. We set a maximum length of 512 tokens, and any extra tokens are truncated. If a text is shorter than the maximum length, we add padding tokens (typically zeros) to make it fit the required size. Padding ensures that all inputs have the same length, which is essential for the model to process them efficiently in batches. The processed inputs are returned as PyTorch tensors using `return_tensors="pt"` to ensure compatibility with the model's architecture. On the decoder side, the normalized targets are tokenized with a smaller maximum length of approximately 128 tokens, reflecting the typically shorter nature of normalized claims. Padding tokens in the target sequence are assigned the value `labels=-100` to prevent them from affecting the loss computation.

5.2. Model Training

Model training is conducted using Hugging Face's `Seq2SeqTrainer`[4] framework, which abstracts much of the training loop while allowing customization. The training setup uses a learning rate of $3e-5$ and a batch size of 4, optimized for limited GPU memory. We train the model for 5 epochs. After each epoch, the model is evaluated using METEOR[20] score[5]. To improve efficiency and reduce memory usage, training is performed in mixed-precision (FP16). Additionally, `gradient_accumulation_steps` is used to simulate a larger batch size, improving optimization stability.

5.3. Decoding and Inference

When the model is tested, it generated the output step by step, using each previously generated token as context for the next one. This technique, known as autoregressive decoding[22], allowed the model to generate one token at a time. After generating the first token, the model incorporated it into the context to produce the next token, continuing this process until the full sequence is created. To improve

[3]https://huggingface.co/facebook/bart-large

[4]https://www.kaggle.com/code/simonepiocaronia/seq2seq-trainer-transfer-learning

[5]https://huggingface.co/spaces/evaluate-metric/meteor

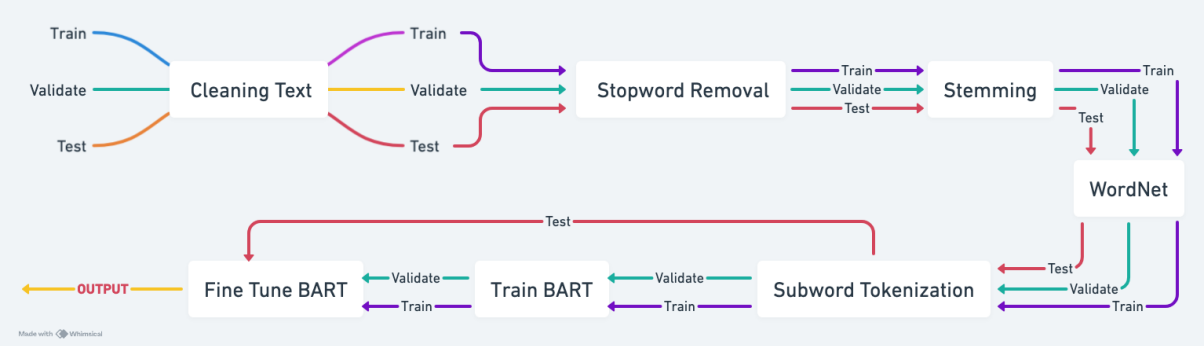

Figure 2: Claim Extraction and Normalization Workflow.[21]

the quality of generated sequences, beam search with `num_beams=4` is employed. This method enabled the model to explore multiple candidate sequences and select the best one. Once decoding is completed, the outputs are post-processed using `batch_decode(..., skip_special_tokens=True)` to remove special tokens such as `<s>` and `</s>`, resulting in clean, human-readable normalized claims.

5.4. Loss Function and Evaluation

The *Cross-entropy loss*[23] was used as the loss function. To prevent padding tokens from influencing the loss computation, we mask them using the `labels=-100` setting. Evaluation is based on sequence-level metrics, particularly **METEOR**, which measured both the fluency and fidelity of the normalized outputs. Overall, this four-module methodology took the advantage of BART's powerful pre-trained capabilities and fine-tuned them through a structured training and decoding pipeline. By approaching the problem as a generation task, our system successfully transforms noisy, informal claim statements into clear, normalized outputs with high semantic alignment.

Algorithm 1 Text Normalization with BART Large

Require: Dataset $D_{\text{train}}, D_{\text{val}}, D_{\text{test}}$, model M, tokenizer τ
Ensure: Predictions Y_{test} saved to CSV

Load and clean CSVs ▷ Remove nulls, rename columns
Initialize tokenizer τ and model $M \leftarrow$ BART Large
for each sample $x \in D_{\text{train}} \cup D_{\text{val}}$ **do**
    Format as: `"normalize claim: "` + `input_text` ▷ Prefix task instruction
    Tokenize input and target text ▷ Max length, truncation, padding
end for
Convert to HuggingFace Dataset objects ▷ Use 'Dataset.from$_p$*andas*`
Set up training arguments ▷ Batch size, learning rate, epochs, evaluation strategy
Initialize Trainer ▷ Pass model, args, datasets, and data collator
Train the model $\rightarrow M_{\text{best}}$ ▷ Best checkpoint saved based on eval metric
Evaluation:
Add prefix to test inputs ▷ Prepare test examples like training
Generate predictions $Y_{\text{test}} \leftarrow M_{\text{best}}$ ▷ Use 'model.generate()'
Save predictions to CSV file ▷ Output results to local storage

6. Observation

As shown in Table 2, there is significant variation in the performance of normalization models across social media posts. While the Gold Labels maintain factual and contextually accurate claims, the

outputs from models like Pegasus and T5 Small often deviate from the original intent or introduce hallucinations. BART Large and its fine-tuned variant tend to produce more coherent summaries, though accuracy still varies. Notably, the fine-tuned BART model shows improvement in structure but occasionally misinterprets the core content. These inconsistencies highlight the challenge of reliable claim normalization, especially in complex or repetitive posts.

Table 2
Table of Social Media Posts and their Corresponding Model Normalizations.

Posts and their respective normalized form
Post: Though the US is not part of the ICC, if the alleged crime happened in an ICC member country, the Americans could still be investigated under the jurisdiction of the ICC. In 2019, the ICC rejected investigating the alleged war crimes of Americans in Afghanistan because, the judges say, "it would not serve the interests of peace." At the ICC, different rules apply for the powerful. *ICC rejects request to investigate war crimes in Afghanistan — 12 April 2019, REUTERS:* Judges say Afghanistan's current situation made the prospect for a successful investigation "limited". The ICC has rejected its prosecutor's request to investigate alleged war crimes in Afghanistan. **Model Outputs:** • **Gold Label:** Report shows ICC practices selective justice in Philippine drug war probe • **T5 Small:** the u is not part of the icc if the alleged crime happened in an icc • **Pegasus:** a facebook post claim that the u president joe biden said that the donald trump is not be the u • **BART Base:** 2019 icc reject investig alleg war crime american afghanistan judg say would serv interest peac • **BART Large:** the icc rejected investigating the covid19 vaccine programme in 2019
Post: Pence unfollowed Trump, and then changed his banner picture to Biden and Kamala. He's outta there. Mike Pence - @MikePence Vice President of the United States Joined February 2009 | 48 Following | 6.1M Followers **Model Outputs:** • **Gold Label:** Vice President Mike Pence unfollowed the president and changed his Twitter banner to an image of Joe Biden and Kamala Harris • **T5 Small:** joe biden changed his facebook banner to joe biden and kamala harris • **Pegasus:** say joe biden and donald trump • **BART Base:** say penc chang banner pictur biden kamala harri • **BART Large:** pennsylvania senator changed his twitter banner to biden and kamala harris

7. Results and Discussion

In this section, we present the empirical evaluation of different transformer-based models on the claim normalization task. The objective is to assess how well each model can transform noisy, informal inputs into concise and verifiable claims. The comparison is based on METEOR, a metric suitable for evaluating the difference between predicted and reference outputs. We divide the discussion into two parts: quantitative results and model-wise observations.

7.1. Results

We evaluated the performance of several transformer-based models on the task of claim normalization, including **T5 Small**[6], **Pegasus**[7], and both **BART Base** [8] and **BART Large**[9]. The task involved transforming informal and often noisy social media posts into clearly structured, verifiable claims.

Table 3
Comparison of METEOR scores between different models.

Serial No	Model Used	METEOR
1	T5 Small	0.2539450609
2	Pegasus	0.0558825213
3	BART Base	0.2962715775
4	BART Large	0.3098173489

The experimental results are presented in Table 3. Among the tested models, **Pegasus** performed the weakest with a METEOR of 0.0559. While Pegasus has demonstrated strength in summarization tasks, it underperformed in this task due to the structural differences between summarization and claim normalization.

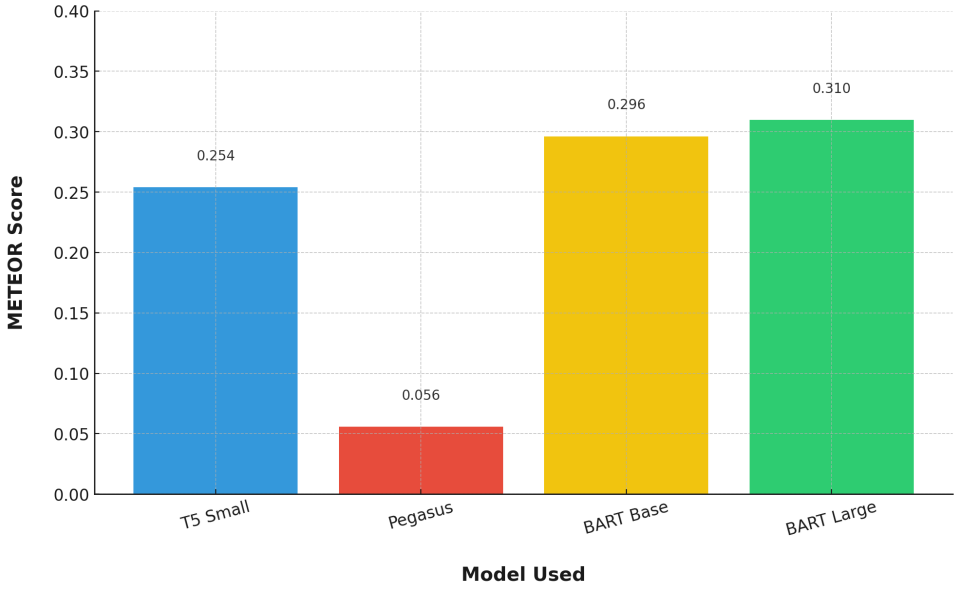

Figure 3: Bar plot comparing the METEOR scores between different models.

T5 Small achieved a better METEOR of 0.2539, indicating moderate effectiveness despite its smaller architecture. **BART Base** further improved upon this with a score of 0.2962. However, the highest performance was recorded by **BART Large**, which achieved the best METEOR of 0.3098, outperforming all other models.

7.2. Discussion

The superior performance of **BART Large** highlights the importance of both model capacity and domain-specific training. Fine-tuning the model on the task-specific dataset enabled it to learn the syntactic and semantic patterns characteristic of informal social media claims. Factors such as longer training

[6]https://huggingface.co/google-t5/t5-base
[7]https://huggingface.co/google/pegasus-large
[8]https://huggingface.co/facebook/bart-large
[9]https://huggingface.co/facebook/bart-large

duration (10 epochs), larger batch size (16), and beam search decoding (`num_beams=8`) contributed to generating more coherent and accurate outputs. Additionally, the use of a decoding temperature of 0.7 helped reduce randomness and improved the consistency of generated sequences.

In contrast, **Pegasus**, although optimized for abstractive summarization, failed to generalize to the claim normalization task. Despite incorporating preprocessing steps such as lemmatization and text cleaning, its outputs often lacked specificity and structure. **T5 Small**, while showing better performance than Pegasus, struggled with complex sentence transformations due to its limited model capacity. T5 Small is a smaller model, which impacts its ability to effectively process and normalize more complex or noisy claims that require nuanced transformations.

Overall, the results emphasize that both model scale and targeted fine-tuning are critical for high-quality claim normalization. Larger transformer models, such as BART-Large, are more capable of handling the complex structure and varied linguistic nuances present in social media content, especially when fine-tuned on task-specific data. This shows that both the architecture's depth and the quality of task-specific training are crucial to achieving the best performance in claim normalization tasks.

8. Error Analysis

While the fine-tuned BART Large model demonstrated strong performance on the claim normalization task, a closer examination of its predictions reveals several recurring error patterns that highlight important limitations and opportunities for improvement, as explained in Table 4.

Table 4
Example of each error analysis of identification through the BART Large Model on the Claim Normalization Task.

Error Category	Post	Normalized Claim	Issue/Explanation
Over-normalization	Though the US is not part of the ICC, if the alleged crime happened in an ICC member country...	2019 icc reject investig alleg war crime american afghanistan judg say would serv interest peac	The model removed the important qualifiers "alleged" and "could," simplifying the claim to a definitive statement and distorting the original meaning.
Sarcasm	Something to #consider... 40 years worth of research, no vaccine for HIV...	40 year worth research covid19 vaccin hiv least 100 year researchno vaccin cancer ongo research vaccin common cold less	The sarcastic tone and rhetorical questioning were lost in the normalization process, turning the sarcastic claim into a straightforward statement.
Negation Handling	Though the US is not part of the ICC, if the alleged crime happened in an ICC member country...	ICC rejects request to investigate war crimes in Afghanistan	The model lost the negation "not part of the ICC," which altered the intended meaning of the claim, implying the opposite.
Multilingual Noise	Legit ba toh?? Legit ba toh?? Legit ba toh?? Legit ba toh?? CSC Republic of the Philippines...	civil servic exam updat complianc govern philippin	The model struggled to handle multilingual noise, resulting in loss of cultural context and intent.

The errors encountered includes:

1. **Over-normalization:** One common error observed was over-normalization, where the model simplified claims to the point of losing important nuances. Speculative phrases like *"could"* or *"may"* were often removed, turning tentative claims into definitive ones. This is particularly

problematic in areas such as politics and public health, where qualifiers are crucial. For instance, "The US may investigate the alleged war crimes" was normalized to "The US investigates war crimes," changing the original meaning.

2. **Sarcasm:** The model frequently misinterprets sarcastic or rhetorical posts as literal, distorting the original tone. For example, the sarcastic statement *"Vaccines for HIV? That's totally going to work!"* was normalized as *"Vaccines for HIV will work."* This misinterpretation, especially in health-related content, can lead to misleading conclusions. Future improvements could incorporate sarcasm detection models, such as sentiment analysis or contextual reasoning techniques, to preserve the intended tone.
3. **Negation Handling:** Negation handling is another critical issue. The model often loses or misinterprets negations, leading to claims with the opposite meaning. For instance, *"The US is not a part of the ICC"* was normalized as *"The US is a part of the ICC"*—altering the original message. To address this, integrating a negation detection module could preserve the integrity of negated statements.
4. **Complex Sentence Handling:** The model tends to simplify complex sentences, resulting in partial or under-specified normalization. While this improves readability, important context can be lost. For instance, multi-clause sentences like *"Though the US is not part of the ICC, it could still be investigated if the crime happened in an ICC member country"* were simplified too much. Incorporating a dependency parsing layer or a sentence simplification algorithm can help retain complex structures while ensuring clarity.
5. **Multilingual Noise:** The multilingual nature of social media content poses significant challenges. The model struggles to preserve context and cultural nuances, particularly in posts with slang or mixed languages. For example, the Filipino phrase *"Legit ba toh??"* loses its cultural meaning when normalized. Using multilingual transformers such as mBERT or XLM-R could improve performance by capturing linguistic diversity and contextual subtleties in multi-language content.

These error patterns highlight key areas for improvement. Specifically, enhancing negation handling, sarcasm detection, multilingual robustness, and preserving contextual nuances could significantly increase the accuracy and trustworthiness of claim normalization models. These improvements will help adapt the system for real-world applications like automated fact-checking, where precision and contextual integrity are crucial.

9. Conclusion

This paper addresses the critical issue of misinformation detection on social media by focusing on the Claim Extraction and Normalization task at CLEF 2025. We propose a novel approach that leverages a fine-tuned BART-Large transformer model to automatically extract and normalize factual claims from noisy, user-generated content. Our method outperforms other transformer-based models, including T5-Small and Pegasus, achieving the highest METEOR score of 0.3098, demonstrating its superior performance in transforming informal social media posts into clear, concise, and verifiable claims.

The results underscore the importance of model scale and task-specific fine-tuning in achieving high-quality claim normalization. Additionally, our methodology, which involves careful data preprocessing, mixed-precision optimization, and beam search decoding, significantly improves the fluency and fidelity of the normalized claims, making them suitable for downstream fact-checking systems.

Despite these successes, several challenges remain, particularly in handling sarcasm, negation, and the multilingual noise inherent in social media data. Error analysis revealed issues such as over-normalization, where critical qualifiers were lost, and difficulties in handling culturally specific expressions. These limitations highlight the need for future improvements, such as incorporating better contextual understanding, improved handling of negation, and enhanced multilingual robustness.

Beyond these technical challenges, it is crucial to consider the ethical implications of deploying automated misinformation detection systems. Automated fact-checking systems must be accurate and

transparent, ensuring that their outputs are both reliable and verifiable. Misleading or incorrect outputs could have significant social consequences, especially in politically sensitive contexts or public health matters. Therefore, developers must prioritize transparency, human oversight, and accountability in future iterations of these systems.

10. Future Works

Future work on this task will focus on addressing the limitations identified in the current study and expanding the scope of claim extraction and normalization. Some key directions for future research include:

- **Handling Sarcasm and Negation More Effectively:** One of the key challenges identified was the model's difficulty in handling sarcasm and negation. Further work could involve incorporating specialized sentiment analysis or sarcasm detection modules to help the model distinguish between literal and figurative language, which is crucial in social media posts where sarcasm is often used.
- **Multilingual Robustness:** As social media content is multilingual, it is important to improve the model's ability to handle multilingual noise. Future models could explore the use of multilingual transformers (e.g., mBERT, XLM-R) or domain-specific fine-tuning to handle the linguistic and cultural diversity of online content more effectively.
- **Bias and Fairness:** The reliance on large-scale social media data presents potential bias and fairness concerns. Future work should focus on mitigating biases in the training data to prevent the model from amplifying social, political, or cultural biases. Techniques such as fairness constraints during training or expanding the diversity of the dataset can help address these concerns.
- **Real-Time Misinformation Detection:** Real-time application of claim normalization in fact-checking systems remains a challenge. Future efforts should explore ways to make the model more computationally efficient, ensuring it can handle large volumes of data while maintaining high accuracy and speed.
- **Ethical and Social Implications:** As automated systems for fact-checking become more widespread, their ethical implications must be further explored. The role of human oversight in these systems will be essential to ensure that automated decisions are accurate and socially responsible. Additionally, ethical frameworks should be developed to address potential misuse or errors in the system.
- **Cross-Domain Claim Normalization:** While this paper focuses on social media posts, the claim normalization approach could be extended to other domains, such as news articles, blogs, or scientific papers. Future work could explore the effectiveness of the model across diverse types of content and domains to create a more generalized approach for claim extraction and verification.

In summary, addressing these challenges will significantly improve the robustness and scalability of the claim extraction and normalization system. By enhancing multilingual capabilities, tackling ethical concerns, and improving model efficiency, the proposed approach could become a valuable tool in combating misinformation on a global scale. These future directions will pave the way for more accurate, inclusive, and real-time misinformation detection systems that can be deployed across various platforms and domains.

Acknowledgements

We would like to express our sincere gratitude to the organizers of the CheckThat! Lab @ CLEF 2025 for designing such an insightful and timely shared task. Participating in this challenge provided us with a valuable opportunity to explore the evolving boundaries of AI-generated text and its detectability in real-world contexts.

We are especially thankful to the CLEF community for providing robust infrastructure, clearly defined evaluation protocols, and constructive feedback throughout the process. Their dedication to fostering innovation in authorship verification and stylometry continues to inspire meaningful research.

We would also like to acknowledge the support and encouragement from the Department of Computer Science and Engineering, Jadavpur University. Special thanks to our mentors and peers for their valuable discussions, which greatly contributed to the development and refinement of our system.

Finally, we are grateful for the open-source tools and platforms, including Hugging Face Transformers and Python libraries, that made this research accessible and reproducible.

Declaration on Generative AI

During the preparation of this work, the author(s) used OpenAI-GPT-4 in order to: Grammar and spelling check. Further, the author(s) used Whimsical for figures 2 in order to: Generate images. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M. Sundriyal, T. Chakraborty, P. Nakov, From chaos to clarity: Claim normalization to empower fact-checking, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, 2023, pp. 6594–6609. URL: https://doi.org/10.18653/v1/2023.findings-emnlp.439.

[2] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G. Da San Martino, Automated fact-checking for assisting human fact-checkers, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), Survey Track, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 4551–4558. URL: https://www.ijcai.org/proceedings/2021/627, survey Track.

[3] R. Panchendrarajan, A. Zubiaga, Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research, arXiv preprint arXiv:2401.11969v3 (2024). URL: https://arxiv.org/abs/2401.11969, preprint submitted to Natural Language Processing, March 19, 2024.

[4] L. Majer, J. Šnajder, Claim check-worthiness detection: How well do llms grasp annotation guidelines?, in: Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER), Association for Computational Linguistics, 2024, pp. 245–263. URL: https://aclanthology.org/2024.eacl-long.5, accessed: 2024-03-19.

[5] Z. Deng, M. Schlichtkrull, A. Vlachos, Document-level claim extraction and decontextualisation for fact-checking, arXiv preprint arXiv:2406.03239v2 (2024). URL: https://arxiv.org/abs/2406.03239, preprint submitted to Fact Extraction and VERification, June 2024.

[6] 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, 2011, pp. 48–57. URL: https://www.aclweb.org/anthology/W11-0707.

[7] 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.

[8] 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, Association for Computational Linguistics, 2019, pp. 558–563.

[9] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. D. S. Martino, M. Hasanain, R. Suwaileh, F. Haouari, Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media, in: Proceedings of the 42nd European Conference on Information Retrieval, Lisbon, Portugal, 2020, pp. 499–507.

[10] M. Rodríguez-Ibáñez, A. Casaéz-Ventura, F. Castejón-Mateos, P.-M. Cuenca-Jiménez, A review on sentiment analysis from social media platforms, Expert Systems With Applications 223 (2023) 119862. URL: https://doi.org/10.1016/j.eswa.2023.119862. doi:10.1016/j.eswa.2023.119862, available online 14 March 2023.

[11] L. Basyal, M. Sanghvi, Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models, arXiv preprint arXiv:2310.10449v2 (2023). URL: https://arxiv.org/abs/2310.10449, accessed: 2023-10-14.

[12] W. Kryscinski, B. McCann, C. Xiong, R. Socher, Evaluating the factual consistency of abstractive text summarization, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 9332–9346.

[13] P. Utama, J. Bambrick, N. Moosavi, I. Gurevych, Falsesum: Generating document-level nli examples for recognizing factual inconsistency in summarization, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 2763–2776.

[14] A. Fan, D. Grangier, M. Auli, Controllable abstractive summarization, in: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 45–54.

[15] G. S. Cheema, S. Hakimov, A. Sittar, E. Müller-Budack, C. Otto, R. Ewerth, MM-Claims: A dataset for multimodal claim detection in social media, in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 962–979. URL: https://aclanthology.org/2022.findings-naacl.75.

[16] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019). URL: https://arxiv.org/abs/1910.13461, facebook AI.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805v2 (2019). URL: https://arxiv.org/abs/1810.04805v2.

[18] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, arXiv preprint arXiv:1810.04805v2 (2018). URL: https://arxiv.org/abs/1810.04805v2.

[19] L. Kozma, J. Voderholzer, Theoretical analysis of byte-pair encoding, arXiv preprint arXiv:2411.08671v1 (2024). URL: https://arxiv.org/abs/2411.08671v1, supported by DFG Grant KO 6140/1-2.

[20] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://www.aclweb.org/anthology/W05-0909/.

[21] Whimsical Inc., Whimsical – Visual Workspace for Collaboration, 2025. URL: https://whimsical.com/, accessed: 2025-07-06.

[22] D. Schuurmans, H. Dai, F. Zanini, Autoregressive large language models are computationally universal, arXiv preprint arXiv:2410.03170v1 (2024). URL: https://arxiv.org/abs/2410.03170v1.

[23] K. Krishna, J. Sedoc, G. Neubig, Y. Tsvetkov, Automatic detection of machine-generated text: A critical survey, arXiv preprint arXiv:2304.07288 (2023).