Investigators at CheckThat! 2025: Using LLMs to Improve Fact-Checking*

Notebook for the CheckThat! Lab at CLEF 2025

Syed Muhammad Ather Hashmi[1,*,†], Sidra Aamir[1,†], Muhammad Anas[1,†], Turab Usmani[1,†], Faisal Alvi[1] and Abdul Samad[1]

[1]Department of Computer Science, Dhanani School of Science and Engineering, Habib University, Karachi, Pakistan

Abstract

This paper presents our approach to addressing two critical challenges in the CheckThat! Lab [1][2] tasks using Large Language Models (LLMs). For Task 1, we developed both monolingual and multilingual models to classify texts as subjective or objective, a crucial step in identifying claims that may require fact-checking. For Task 2, we implemented multiple LLM-based approaches for claim matching across different languages, including the use of FLAN-T5, BART-base, and LLaMA-3-2-1B models. We evaluated our methods on multilingual and monolingual datasets including English, Spanish, Korean, Italian, German, Arabic, and Bulgarian. Our results demonstrate the effectiveness of fine-tuned LLMs for these tasks, particularly highlighting the benefits of transfer learning and multilingual capabilities in low-resource settings. Furthermore, we explore zero-shot and few-shot learning capabilities of recent models, offering insights into the potential of LLMs for automated fact-checking pipelines. For Task 1, mDeBERTAv3-subjectivity-multilingual model delivered the best results, whereas for Task 2, google/flan-t5-base yielded the optimal results on the dev-set.

Keywords

CLEF CheckThat!, fact-checking, transformer models, binary classification, encoder-decoder, subjectivity classification, claim normalization, natural language processing, social media post normalisation

1. Introduction

With the rise of misinformation across digital platforms, fact-checking has become more critical than ever. However, automating this process effectively introduces a range of challenges. Two essential components in the automation pipeline are subjectivity detection (SD) and claim normalization, both of which play a vital role in making fact-checking more reliable and scalable.

Our work focuses on two complementary tasks from the CheckThat! Lab:

1. **Subjectivity Classification (Task 1)**: Identifying whether a piece of text is subjective (expressing opinions or personal views) or objective (presenting factual information), serving as a crucial filtering step in fact-checking pipelines [3].
2. **Claim Normalization (Task 2)**: Simplifying noisy, unstructured social media posts into concise, factual claims. These posts may be written in any of the 20 languages specified in the task [4].

Task 1 encompasses three settings: monolingual, multilingual, and zero-shot prediction. The monolingual and multilingual settings are offered in five languages—Arabic, Bulgarian, English, German, and Italian.

The key challenge lies in the subjective nature of "subjectivity" itself, which can vary significantly based on linguistic and cultural contexts. Subjective content often conveys bias and opinion, which can affect claim extraction, normalization, and overall bias detection.

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

*Corresponding author.

†These authors contributed equally.

sh07554@st.habib.edu.pk (S. M. A. Hashmi); sa07316@st.habib.edu.pk (S. Aamir); ma08458@st.habib.edu.pk (M. Anas); tu08125@st.habib.edu.pk (T. Usmani); faisal.alvi@sse.habib.edu.pk (F. Alvi); abdul.samad@sse.habib.edu.pk (A. Samad)

0009-0009-1710-8230 (S. M. A. Hashmi); 0009-0004-5659-2268 (S. Aamir); 0009-0002-9064-1567 (M. Anas); 0009-0009-1374-5144 (T. Usmani); 0000-0003-3827-7710 (F. Alvi); 0009-0009-5166-6412 (A. Samad)

To address this, we fine-tuned transformer-based models such as `mdeberta-v3` and `twitter-roberta-base-sentiment` [5].

Task 2, **Claim Extraction and Normalization**, aims to transform informal social media posts into clear, checkable claims. This assists fact-checkers by stripping away irrelevant information and highlighting the core factual elements. For this task, we fine-tuned advanced models such as LLaMA-3.2 [6], BART [7], and T5[8].

Both tasks share a common objective: to enhance the accuracy and efficiency of fact-checking by improving the reliability of information extraction. The findings from our experiments contribute to building more effective misinformation detection systems and improving the trustworthiness of automated fact-checking tools.

2. Literature Review

2.1. Task 1

Subjectivity detection in sentences has been studied for years, particularly in sentiment analysis and bias detection. Earlier methods relied on lexicon-based approaches and statistical models,but researchers have found that fine-tuning models for specific languages leads to better accuracy, and data augmentation techniques, such as GPT-3-generated samples, help balance class distributions and improve model robustness.[9]

CheckThat! 2024 task 2 [10] evaluated subjectivity detection models across five languages: English, Arabic, Italian, Bulgarian, and German, along with a multilingual setting. The baseline model was a multilingual **SentenceBERT** with a logistic regression classifier. Competing teams applied various strategies, including fine-tuning models, regression techniques, k-nearest neighbours, and support vector machines. The most effective approach involved fine-tuning a BERT classifier pretrained for sentiment analysis in each language. While these methods significantly improved performance over baseline models, but challenges remained for certain languages like Arabic and Bulgarian [10]

Building on these developments, [11] examines how large language models (LLMs) detect subjectivity in news articles. Traditional methods, including lexicon-based and machine learning models like SVMs lacked generalizability, while models such as BERT and RoBERTa improved contextual understanding. Recently, LLMs like GPT-3.5, GPT-4, and Gemini have been used in-context learning (ICL) for classification, though performance depends on prompt quality. The study [11] also finds that fine-tuning achieves high accuracy on in-distribution data but struggles with out-of-distribution (OOD) generalization. Zero-shot and few-shot ICL methods perform well with the optimized prompts, particularly Chain-of-Thought (CoT) prompting.

An approach [12] for subjectivity testing of the news article sentences, by augmenting a small dataset using NLTK and WordNet i.e. by replacing random words in each sentence with their synonym. Initial modelling was done with mDeBERTA, achieving an F1 score of 0.76. Later, the dataset was re-augmented using the advanced Google Gemini Model, creating a balanced dataset with three similar sentences for the label Objective and five similar sentences for the label Subjective. Different models and ensemble techniques were applied to RoBERTa-base alone, resulting in MACRO F1 score of 0.708 and SUBJ F1 of 0.54. The results suggested that a low SUBJ F1 may be either due to the data augmentation method or noise and less distinctive features in the SUBJ class of data. [12]

2.2. Task 2

Claim normalization is a relatively new concept in CheckThat Labs but shares similarities with claim detection and works to extract and simplify the central claim in the given social media posts.[13]. It has traditionally been approached as a semantic textual similarity task, with early systems using TF-IDF and BM25 retrieval methods [14]. Recent advances have leveraged dense retrieval approaches with dual-encoder architectures such as Sentence-BERT [15]. Transformer based models like T5 and BART have been effective at condensing complex text, but they often struggle with maintaining factual consistency.

The CACN framework builds on these existing methods by introducing reverse check-worthiness, ensuring that only factually relevant information is retained.[13]

The paper [13] also found that in-context learning with GPT-3 was surprisingly effective, often outperforming fine-tuned models when given well-crafted prompts. Both subjectivity detection and claim normalization contribute to the broader goal of automated misinformation detection. Recent trends highlight the growing importance of multilingual approaches and context-aware reasoning, as well as the need for robust, accurate, fair and generalizable evaluation metrics to assess performance in text simplification and fact-checking tasks. [16] [17]

3. Methodology

3.1. Task1

The task was offered in three settings: monolingual, multilingual and zero-shot.

3.1.1. Dataset

The subjectivity classification task utilize a multilingual dataset comprising five languages: English, Italian, German, Bulgarian, and Arabic. Each language dataset is structured in TSV (Tab-Separated Values) format with three columns: sentence_id, sentence, and label. The labels are binary, with "OBJ" indicating objective text and "SUBJ" indicating subjective text.

A notable characteristic across all languages is the class imbalance, with objective sentences generally outnumbering subjective ones. This imbalance varies by language, with Italian showing the highest proportion of objective content (76% in the training set) and Bulgarian having the most balanced distribution.

To address the class imbalance and improve model performance, we implemented data augmentation techniques for English, Arabic, and Bulgarian languages using the Gemini 2.0 model, generating an additional 1,000 samples for each language.

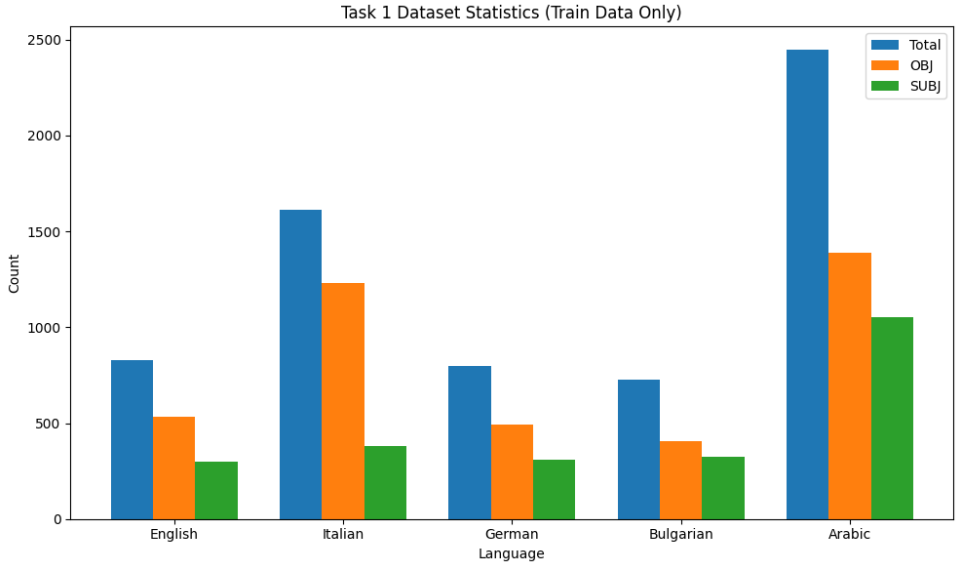


Figure 1: Statistics about the training data

This augmentation significantly improved model performance, particularly for languages with smaller training sets.

3.1.2. Monolingual Dataset setting

For every language other than English, the general approach was to fine-tune a transformer-based model. We tried several models, including BERT, RoBERTa, DeBERTa, also models that had already been fine-tuned on a specific language, like CAMeLBERT for Arabic.
For English, we used focal loss with class weights because the model was struggling with subjectivity. Focal loss specifically addresses this by reducing the impact of easily classified objective instances, forcing the model to emphasise challenging subjective examples. Class weights were applied to further emphasise the subjective class during training. This combined approach aimed to enhance the model's ability to accurately identify and predict subjective text.

Data Augmentation was done to increase the size of the dataset, add more subjective entries so that the models could better predict them.

The table shows the different models we used and the respective hyperparameters used for fine-tuning. We kept weight decay, batch size and many other parameters the same for every language.

Table 1
Models and hyperparameters used for each language.

Model	Epochs	Learning Rate
twitter-roberta-base-sentiment	5	2×10^{-5}
bert-base-arabic-camelbert-mix	3	1×10^{-5}
bert-base-italian-cased-sentiment	5	2×10^{-5}
XLM-RoBERTa-German-sentiment	3	2×10^{-5}
xnli3.0_bulgarian_model	3	2×10^{-5}

The evaluation of our trained model on the test-dev dataset proceeded by first tokenizing the input data. These tokens were then passed through the model. The resulting output logits were subsequently transformed into probability distributions. Classification into "Subj" or "Obj" categories was achieved via a Sigmoid activation followed by an Argmax operation. The resulting predicted labels("Subj"/"Obj") were compared to the gold standard labels provided by CheckThat Lab [18] for performance assessment.

3.1.3. Multilingual & Zero-shot Dataset settings

For the Multilingual setting, we first finetuned different models, including the bert-base-multilingual-uncased-sentiment model and got the best results on finetuning mdebertav3-subjectivity-multilingual, which is finetuned on subjectivity classification tasks across multiple languages. [19]

Since the dataset exhibited class imbalance, with dominance of objective labels. To mitigate this imbalance, we generated around 1000 samples with an equal number of subjective and objective labels, coupled with the given dataset to make the total training data of 8926 entries and validation data of 2393 entries. The samples were generated through Google AI Studio using the Gemini-2.0 model [20] using a few-shot prompt containing some random examples from training data.

The learning rate was set to **3e-5** and the model was run for 8 epochs. After training, we evaluated our model on the test-dev dataset after first tokenizing the input data. These tokens were then passed through the model on languages given in the training data and also on unseen data for Zero-shot setting

3.2. Task 2

The task is structured into two distinct settings: Monolingual and Zero-shot.

3.2.1. Dataset

This task is a generation task offered in 20 languages: English, Arabic, Bengali, Czech, German, Greek, French, Hindi, Korean, Marathi, Indonesian, Dutch, Punjabi, Polish, Portuguese, Romanian, Spanish, Tamil, Telugu, Thai. The bar graph below illustrates the dataset size for each language.

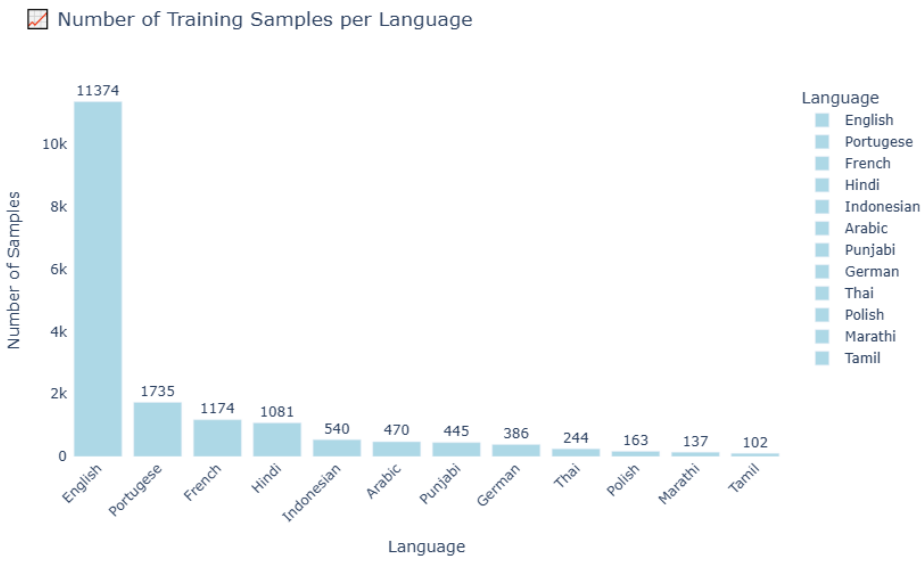


Figure 2: Distribution of dataset sizes across different languages for Task 2 claim normalization.

Dataset was provided in .csv (comma-separated files) for each language, with two columns *post* and *normalized claim*. The task was "given an unstructured, noisy post, write the claim of the post in a normalized and summarized manner".

3.2.2. Monolingual Dataset Setting

In this configuration, the training, validation, and test datasets are all specific to a single language. The model is trained, validated, and tested exclusively on data from that language, allowing it to learn language-specific patterns and structures. The languages included in this setup are English, German, French, Spanish, Portuguese, Hindi, Marathi, Punjabi, Tamil, Arabic, Thai, Indonesian, and Polish.

We focused on English and Spanish for this section, exploring and experimenting with various models to assess their performance.

Table 2
Model configurations for english.

Model	Epochs	Learning Rate
google/pegasus-xsum	5	2×10^{-5}
facebook/bart-base	5	2×10^{-5}
google/flan-t5-base	5	2×10^{-5}
meta/llama-3.2-1b	5	2×10^{-5}

For English, a huge dataset was provided. We fine-tuned the training set using the parameters provided in Table 2. The summarisation and generation models that were used are. Pegasus-xsum, Bart-base, Flan-t5-base and quantized Llama 3.2 1B.

Before finetuning, a pre-processing pipeline was set up. It was noted that the training set had some records in other languages; hence, these records were filtered out from the model training process, in order to allow the training process to gain more insightful patterns from the particular monolingual language dataset. Training arguments were kept constant throughout, changing the models only to decide on a decisive model.

Table 3
Model configurations for spanish.

Model	Epochs	Learning Rate
facebook/bart-base	20	5×10^{-5}
facebook/bart-large	20	5×10^{-5}
google/flan-t5-base	20	5×10^{-5}
google/mt5-base	20	5×10^{-5}
google/flan-t5-base	5	2×10^{-5}

3.2.3. Zero-Shot setting

In this setting, only the test dataset is available for the target language, with no corresponding training or validation data. The participants were allowed to utilize training data from other languages or conduct a zero-shot evaluation using LLMs, testing the model on the target language without prior exposure to its data. This setup assesses the model's ability to generalize to unseen languages. The languages in this category are Dutch, Romanian, Bengali, Telugu, Korean, Greek, and Czech.

We opted for zero-shot learning on Korean, adopting an approach in which the model was initially trained on Spanish data and subsequently evaluated using the Korean test dataset. This design choice was taken to test the model's capacity and performance on zero-shot cases, when a different Western language Family (Romance) was used to test on a Koreanic, eastern language family member. This attempt was to conduct a test for cross-lingual transfer, evaluating the model's ability to generalize to unseen data. The predictions were then submitted to the platform to obtain the performance scores.

4. Results

4.1. Task 1

For the monolingual setting, we finetuned different models for each language to achieve strong performance across multiple languages. The models included RoBERTa, bert-base and XLM-RoBERTa. We achieved F1-scores ranging from 0.74 to 0.79, except for the Arabic language.

Table 4
Macro-average F1 scores and rankings for monolingual setting based on the Leaderboard

Language	Dev-set	Test-set	Rank
English	0.73	0.75	5/24
Arabic	0.54	0.59	3/15
Italian	0.761	0.75	5/15
German	0.789	0.76	6/17
Bulgarian	0.79	-	-

For multi-lingual setting, we selected **mDeBERTa** model, which gave the best results, after experimenting with Bert-base model. Our approach demonstrated impressive cross-lingual transfer capabilities, particularly for languages with limited training data. We made predictions on dev-dataset for the given five languages and recorded the F1 scores below.

Moreover, we also tested our model on a given unlabelled multilingual dataset and unseen languages like Polish, Ukrainian, Greek, and Romanian. Table 6 demonstrates the results of the leaderboard:

Table 5
Multilingual model testing metrics on evaluation dataset.

Language	Macro-F1	Macro-P	Macro-R	Accuracy
Arabic	0.522	0.528	0.528	0.523
Bulgarian	0.742	0.743	0.742	0.748
German	0.852	0.842	0.872	0.866
English	0.728	0.723	0.734	0.789
Italian	0.759	0.775	0.747	0.816

Table 6
Leaderboard results on multilingual and zero-shot settings

Language	F1 Score	Setting	Rank
Zero-shot Languages			
Polish	0.61	Zero-shot	6/15
Ukrainian	0.64	Zero-shot	2/15
Greek	0.45	Zero-shot	6/15
Romanian	0.71	Zero-shot	12/15
Multilingual	0.63	Multilingual	15/17

4.2. Task 2

We evaluated the model performance using two metrics: METEOR and BERTScore. METEOR (Metric for Evaluation of Translation with Explicit Ordering) measures the similarity between generated and reference texts, considering precision, recall, and stemming. BERTScore uses contextual embeddings from BERT to assess semantic similarity, capturing deeper contextual relationships in the text. The BERTScore was used as an optional evaluation metric and is not tested for every language set.

Monolingual Results

Results on English

Table 7
Model-Wise results for English on dev-set(monolingual).

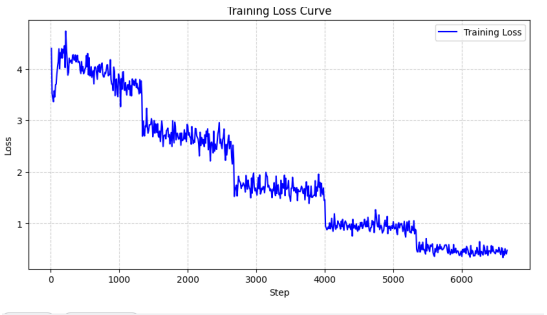
Model	METEOR
google/pegasus-xsum	0.2683
meta/llama-3.2-1b	0.318
google/flan-t5-base	0.418
facebook/bart-base	0.453

Table 7 shows that Bart-Base performed tremendously on the dev-set. However, when the predictions for the test set were made, it was noticed that the Bart output was picking up correct terms from the noisy posts, which were needed in the summary, but it was not able to form a coherent, sensible sentence. Hence, the second-best model, Flan-t5-base was chosen as our best model.[21] Predictions generated through flan-t5 captured the semantics and context of the information.

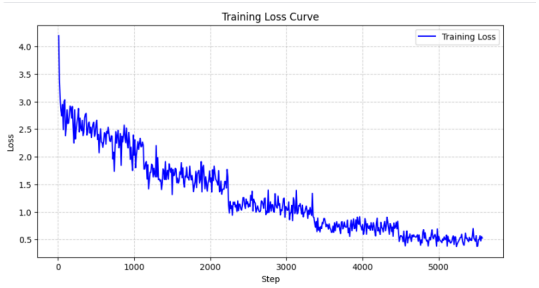
Figure 3 shows the performance of various models while fine-tuning them on the English dataset.

Results on Spanish

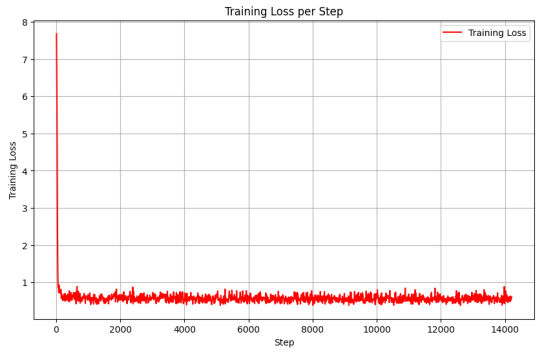
Although Bart showed a great result over the dev-set as shown in Table 8, it generated incoherent predictions on the test set. Hence, the Flan-t5 was chosen next. On epochs=20, it was observed that the



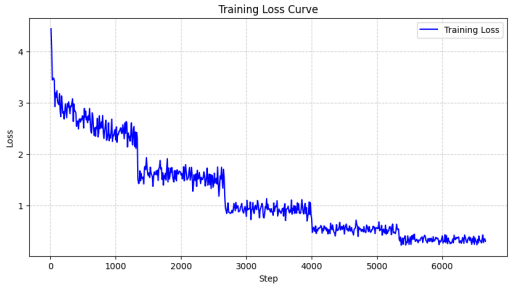
(a) Bart-Base loss curve



(b) Flan-t5 loss curve



(c) Quantized Llama 3.2 1B loss curve



(d) Pegasus-xsum loss curve

Figure 3: Training Loss Curves of Various Models on English dataset

Table 8
Model-Wise results for Spanish (monolingual)

Model	METEOR	BERTScore
facebook/bart-large	0.4326	0.78
facebook/bart-base	0.3125	0.7712
google/flan-t5-base(20epochs)	0.2897	0.7542
google/flan-t5-base(5epochs)	0.313	–
google/mt5-base	0.2437	0.7345

results had been over-fitted, hence a new finetuning was done on epochs=5, details shown in Table 3. The METEOR scored improved significantly. BERTScore was not calculated due to it being not required in the competition and due to shortage of time.

As no dev-set was provided for Korean, nothing could be said about its performance decisively on the dev-set.

Table 9
METEOR Score on Test dataset and Leaderboard Ranking

Language	METEOR	Rank
English	0.3565	10/17
Spanish	0.3447	9/10
Korean	0.0149	5/6

Table 9 shows the results of our models in their respective languages on the test dataset. English and Spanish performed well due to being fine-tuned on the train-dataset, however, the results on Korean were not too appealing.

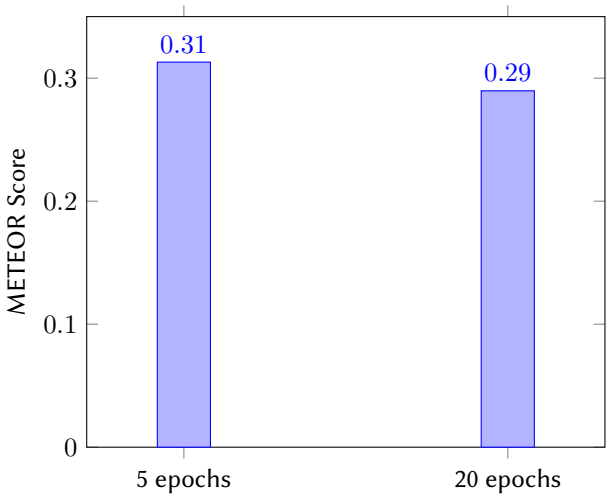


Figure 4: Effect of training epochs on METEOR score for `google/flan-t5-base` on Spanish dataset

5. Discussion

In Task 1, we found a general trend of models predicting objectivity with ease but struggling with subjectivity. Increasing more samples of subjective sentences did not greatly change this. We also found that models tended to overfit the dataset, likely because the models were large in size, and our dataset was small. When we used smaller models (less than 20 million parameters) they did not perform as well. The difficulty in classifying subjectivity, even with data augmentation, points to a core challenge where models likely confuse implicit and explicit expressions of opinion. They may perform well on sentences with clear markers like "I believe," but falter when subjectivity is conveyed subtly through framing, tone, or sarcasm. A significant source of error is likely statements that mix objective facts with subjective assessments, making the overall intent difficult for a model to classify without deeper contextual understanding or world knowledge.

For Task 2, it was observed that summarisation and generation models struggled to perform cleaning the noisy, informal social media posts, filled with emojis and hashtags on the dev-set. However, in both cases, google/flan-t5-base took the lead in generating coherent, semantically and grammatically correct normalized claims, containing the fact of the post comparatively. The poor performance in the zero-shot setting for Task 2, especially for Korean, underscores the difficulties of applying models to languages without specific training data. Failure cases in this context likely stem from several key issues. The model may resort to an overly literal interpretation of the input, failing to grasp the idiomatic expressions, slang, and cultural nuances prevalent in social media posts. This leads to normalized claims that miss the original intent. Furthermore, a lack of language-specific context can cause the model to lose crucial information, misinterpret rhetorical devices, or even "hallucinate" facts not present in the source text. In the most severe cases, the model may produce incoherent or grammatically incorrect output.

6. Conclusion

In conclusion, our research on the CheckThat! Lab tasks demonstrates the use of the capabilities of transformer-based models for multilingual fact-checking applications. For Task 1 (subjectivity classification), our monolingual approach utilizing different models for each language with language-specific tokenizers achieved robust F1-scores ranging from 0.74 to 0.79 except the F1 scores on Arabic language was still 0.53. The multilingual transfer capabilities of the "mdebertav3-subjectivity-multilingual" model were particularly valuable for low-resource languages like Bulgarian and Arabic, moreover generation of more data using Google Gemini improved the F1 score to some extent. Error analysis revealed

specific patterns in model failures, particularly the difficulty in classifying texts containing a mixture of factual information and opinions. Claims requiring world knowledge or containing implicit statements presented particular challenges across all model architectures. For Task 2 (claim normalization), our most successful approach leveraged the google/flan-t5-base model capturing the context and semantic structure of the post and its contents, resulting in the highest METEOR score out of all the other tried models with English Dev-Set having a high record of 0.418 while Spanish had 0.313.

Future work could focus on enhancing the feature representation for subjective class detection and reducing noise through better data preprocessing and augmentation strategies. Moreover implementing more sophisticated data augmentation techniques beyond translation-based methods to improve the accuracy of the model. Ensembling Techniques can be utilised to further enhance the results. Furthermore, an assesment can also be performed using the generative models APIs, and then performing a comparison of the results of gained from fine-tuned models and the API generated responses.

Acknowledgements

The authors would like to acknowledge the support provided by the Office Of Research (OoR) at Habib University, Karachi, Pakistan for funding this project through the internal research grant IRG-2235.

Declaration on Generative AI

During the preparation of this work, the authors employed ChatGPT and Grammarly AI tools for grammar checking, paraphrasing, rewording and consistency checking of sentences. After using the tools, the authors reviewed and edited the content as required and thereby take full responsibility for the publication's content.

References

[1] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. V., The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 467–478.

[2] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. Venktesh, Overview of the CLEF-2025 CheckThat! Lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[3] F. Ruggeri, A. Muti, K. Korre, J. M. Struß, M. Siegel, M. Wiegand, F. Alam, R. Biswas, W. Zaghouani, M. Nawrocka, B. Ivasiuk, G. Razvan, A. Mihail, Overview of the CLEF-2025 CheckThat! lab task 1 on subjectivity in news article, in: [22], 2025.

[4] M. Sundriyal, T. Chakraborty, P. Nakov, Overview of the CLEF-2025 CheckThat! lab task 2 on claim normalization, in: [22], 2025.

[5] F. Barbieri, J. Camacho-Collados, L. Neves, L. Espinosa-Anke, F. Twarog, cardiffnlp/twitter-roberta-base-sentiment, https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment, 2021. Accessed: 2025-05-13.

[6] M. AI, Llama 3.2–1b on hugging face, https://huggingface.co/meta-llama/Llama-3.2-1B, 2024. Accessed: 2025-07-06.

[7] H. Face, Bart model documentation, https://huggingface.co/docs/transformers/en/model_doc/bart, 2024. Accessed: 2025-07-06.

[8] H. Face, T5 model documentation, https://huggingface.co/docs/transformers/en/model_doc/t5, 2024. Accessed: 2025-07-06.

[9] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A structured self-attentive sentence embedding, in: Proc. Int. Conf. Learn. Represent. (ICLR), 2017. URL: https://arxiv.org/abs/1703.03130.

[10] J. M. Struß, F. Ruggeri, A. Barrón-Cedeño, F. Alam, D. Dimitrov, A. Galassi, G. Pachov, I. Koychev, P. Nakov, M. Siegel, M. Wiegand, M. Hasanain, R. Suwaileh, W. Zaghouani, Overview of the clef-2024 checkthat! lab task 2 on subjectivity in news articles, in: Working Notes of CLEF 2024 - Conf. and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proc.*, 2024. URL: https://ceur-ws.org/Vol-3740/paper-25.pdf, notebook for the CheckThat! Lab at CLEF 2024.

[11] M. Shokri, V. Sharma, E. Filatova, S. Jain, S. Levitan, Subjectivity detection in english news using large language models, in: Proc. Workshop Comput. Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), 2024. URL: https://aclanthology.org/2024.wassa-1.17.pdf.

[12] D. Zehra, K. Chandani, M. Khubaib, A. Ali, A. Muhammed, F. Alvi, A. Samad, Checker hacker at checkthat! 2024: Detecting check-worthy claims and analyzing subjectivity with transformers, in: Working Notes of CLEF 2024 - Conf. and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proc.*, 2024. URL: https://ceur-ws.org/Vol-3740/paper-64.pdf, accessed: 2025-03-01.

[13] M. Sundriyal, T. Chakraborty, P. Nakov, From chaos to clarity: Claim normalization to empower fact-checking, in: Findings of the Assoc. for Comput. Linguistics: EMNLP 2023, 2023, pp. 6594–6609. URL: https://aclanthology.org/2023.findings-emnlp.439.pdf.

[14] S. Shaar, A. Nikolov, N. Babulkov, F. Alam, A. Barrón-Cedeño, T. Elsayed, M. Hasanain, R. Suwaileh, F. Haouari, G. Da San Martino, P. Nakov, Overview of checkthat! 2020: Automatic identification and verification of claims in social media, in: Working Notes of CLEF 2020 - Conf. and Labs of the Evaluation Forum, 2020.

[15] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proc. Conf. Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410.

[16] T. Hu, X.-H. Zhou, Unveiling LLM evaluation focused on metrics: challenges and solutions, arXiv (Cornell University) (2024). URL: https://arxiv.org/abs/2404.09135. doi:10.48550/arxiv.2404.09135.

[17] D. Hupkes, M. Giulianelli, V. Dankers, M. Artetxe, Y. Elazar, T. Pimentel, C. Christodoulopoulos, K. Lasri, N. Saphra, A. Sinclair, et al., A taxonomy and review of generalization research in nlp, Nature Machine Intelligence 5 (2023) 1161–1174. doi:10.1038/s42256-023-00747-5.

[18] C. Lab, Checkthat! lab 2025: Task 1 dataset (claim identification), https://gitlab.com/checkthat_lab/clef2025-checkthat-lab/-/tree/main/task1/data, 2025. Accessed: 2025-07-06.

[19] GroNLP, mdebertav3-subjectivity-multilingual, https://huggingface.co/GroNLP/mdebertav3-subjectivity-multilingual, 2023. Accessed: 2025-05-13.

[20] G. DeepMind, Gemini: Our largest and most capable ai models are getting even better, https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/, 2024. Accessed: 2025-07-06.

[21] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, https://arxiv.org/abs/2210.11416, 2022. ArXiv:2210.11416, Creative Commons Attribution 4.0 International.

[22] G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025.