UmuTeam at CheckThat! 2025: Language-Specific versus Multilingual Models for Fact-Checking

Tomás Bernal-Beltrán[1], Ronghao Pan[1], José Antonio García-Díaz[1] and Rafael Valencia-García[1]

[1]*Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain*

Abstract

The rapid spread of disinformation on digital platforms poses a growing threat to public trust and informed decision-making. This challenge is intensified by the high volume of content shared on social media, which complicates timely and accurate fact verification. As a result, there is a pressing need for effective and scalable methods to assess the credibility of online information, in response, the research community has focused on developing automated fact-checking systems capable of operating across diverse languages and linguistic structures. This paper presents our contribution in two subtasks of the CheckThat! Lab at CLEF 2025: *Subjectivity Classification* and *Claim Extraction & Normalization*. For the Subjectivity subtask, we fine-tune encoder-only models using language-specific data in monolingual settings, and a single XLM-RoBERTa model on multilingual data for multilingual and zero-shot scenarios. Results show that language-specific fine-tuning improves accuracy when training data is available, while multilingual training supports cross-lingual generalization. For Claim Extraction & Normalization, we use a generative approach with Flan-T5. We fine-tune one model per language for monolingual settings and a shared model for zero-shot scenarios. Despite lower leaderboard rankings, our system demonstrates the viability of multilingual generative models for structured claim normalization and highlights challenges in low-resource settings.

Keywords

Multilingual Fact Checking, Language-specific models, Textual Analysis, Natural Language Processing

1. Introduction

The proliferation of disinformation across digital platforms poses a significant challenge to contemporary society, undermining public trust and impeding informed decision-making. This issue is exacerbated by the rapid generation of new content on social media, which complicates the timely verification of information. In addition, social media platforms such as Facebook, X, TikTok, YouTube and Instagram, have become the main source of information for a large part of the population, providing instant access to news, opinions and social interactions, facilitating communication and connectivity, but serving as amplifiers for misinformation [1].

There is therefore a critical need for effective mechanisms to assess the veracity of online content, ensuring public access to reliable and trustworthy information. Although manual fact-checking processes are highly accurate, they are becoming increasingly impractical due to the massive volume of content generated every second on social networks [2]. These processes are not only time-consuming and resource-intensive but also face additional complexities in multilingual contexts, where the original claim and the supporting evidence may appear in different languages, further complicating the verification process. This underscores the need for automated, scalable and multilingual fact-checking solutions.

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

tomas.bernalb@um.es (T. Bernal-Beltrán); ronghao.pan@um.es (R. Pan); joseantonio.garcia8@um.es (J. A. García-Díaz); valencia@um.es (R. Valencia-García)

https://portalinvestigacion.um.es/investigadores/2128962/detalle (T. Bernal-Beltrán); https://portalinvestigacion.um.es/investigadores/2149930/detalle (R. Pan); https://portalinvestigacion.um.es/investigadores/332726/detalle (J. A. García-Díaz); https://portalinvestigacion.um.es/investigadores/331575/detalle (R. Valencia-García)

0009-0006-6971-1435 (T. Bernal-Beltrán); 0009-0008-7317-7145 (R. Pan); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)

Recent progress in Natural Language Processing (NLP), particularly through the use of pretrained transformer-based models, has opened up promising avenues for automating the verification of online content [3]. These models have demonstrated impressive capabilities in tasks such as language understanding, text classification, and summarization, all of which are essential components of the fact-checking pipeline. However, despite their potential, deploying such models in real world scenarios remains challenging due to issues such as domain adaptation, data scarcity in low-resource languages, and the risk of model hallucinations. Addressing these limitations requires not only technological innovation but also carefully designed benchmark tasks that simulate realistic fact-checking scenarios across different languages and modalities.

To address these challenges, the research community has been actively working on the development of automated fact-checking systems capable of effectively operate across multiple languages. These systems typically break down the fact-checking process into subtasks that can be partially automated using NLP, such as identifying check-worthy claims, retrieving relevant evidence and assessing veracity [4]. Recent advancements in the field includes leveraging large language models (LLMs) and cross-lingual techniques to enable claim detection and verification in diverse linguistic contexts.

The CheckThat! shared task (CLEF 2025) [5] aims to efficiently identify and normalize (spoil) previously fact-checked claims that correspond to a given social media post in a multilingual context. The overarching goal is to reduce redundancy in fact-checking efforts by detecting which posts need to be verified and linking posts to the most relevant verified claims. It is divided into four "subtasks": (1) **Subjectivity** [6]. Given a text sequence from a news article, the goal is to determine whether it expresses the subjective view of the author behind it or presents an objective view on the covered topic instead; (2) **Claims Extraction & Normalization** [7]. Given a social media post, the goal is to simplify it into a concise form, generating a normalized claim for the given social media post; (3) **Fact-Checking Numerical Claims** [8]. Given a claim, the goal is to verify the numerical quantities and temporal expressions in it, classifying each claim as True, False, or Conflicting based on a short list of evidence; and (4) **Scientific Web Discourse** [9]. This subtask is in turn divided into two "subtasks": (4a) **Scientific Web Discourse Detection** Given a social media post, the goal is to detect if it contains a scientific claim, a reference to a scientific study/publication, or mentions of scientific entities; (4b) **Scientific Claim Source Retrieval**. Given a social media post that implicit references a scientific paper, the goal is to retrieve the mentioned paper from a pool of candidate papers.

For this shared task, we participated in the first two subtasks. For the Subjectivity subtask [6], we propose an approach based on fine-tuning encoder-only models. In the monolingual scenarios, we fine-tuned language-specific encoder-only models, to ensure more specialized language coverage. In multilingual and zero-shot scenarios, we fine-tuned XLM-RoBERTa-Large, a multilingual Transformer-based model, to ensure broader language coverage. For the Claim Extraction & Normalization subtask [7], we propose an approach based on the fine-tuning of Flan-T5-Base [10], a multilingual Text-to-Text Transfer Transformer model, to generate normalized claims from social media post that make a claim.

2. Background Information

The increasing volume of content generated on social networks and the critical need for effective mechanisms to assess the veracity of this content has driven the development of the automation of the fact-checking process. The automated fact-checking process usually consists of several steps, each of which is responsible for performing one of the subtasks that make up the automation of the fact-checking process [4]. Two key steps in this process are: the detection of whether a claim expresses the subjective view of the author behind it or presents an objective view on the covered topic instead (Subjectivity Detection) and the normalization of text to extract the claims present in it in a concise form (Claims Extraction & Normalization). These tasks have been extensively studied in recent literature, both in monolingual and multilingual contexts, with the aim of building robust systems that can operate efficiently across languages and domains, including informal domains such as social networks.

Subjectivity detection (SD), which is understood as the task of classifying text fragments as either

subjective, meaning statements that contain opinions, feelings, or personal appreciations, or objective, meaning statements that are not influenced by opinion, has evolved significantly in recent years. It is considered a sub-problem within sentiment analysis (SA) [11], and has applications in opinion mining, bias detection and fact checking. Early approaches used in SD were based on the use of lexical resources of subjectivity, such as lists of words with subjective polarity, and traditional supervised classifiers, trained with linguistic features such as the count of evaluative adjectives, the presence of personal pronouns, or lexical n-grams. An influential example is that of Pang and Lee (2004) [12], who proposed to automatically extract the subjective parts of a review before determining its polarity. While these approaches enabled important advances, they had significant limitations: they were highly language-dependent, difficult to adapt to other domains, and not robust to phenomena such as semantic ambiguity or figurative language.

In recent years, deep learning approaches have overtaken traditional syntax-based methods. Initially, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) with word embeddings were used to better capture the context of subjectivity in a sentence. However, the real breakthrough came with Transformer-based pre-trained language models, which allow capturing deep, contextualised semantic representations, which is key to detecting subjective nuances even in informal contexts. Particularly with the BERT architecture, which represented the new state of the art. For example, a 2022 study [13] showed that a BERT-based multi-task architecture could simultaneously learn to classify subjectivity and polarity, improving performance on both tasks. Moreover, the need to develop robust systems for multiple languages has driven the development and adoption of multilingual models such as mBERT and XLM-RoBERTa. These models have demonstrated an impressive capacity for knowledge transfer across languages. They can be trained on one set of languages and generalize to others without the need for additional training, while maintaining competitive performance [11]. For instance, in [14], the authors highlighted that combining features across six languages results in robust SD, achieving a 4.90% reduction in accuracy errors compared to using English alone. More recent work [15] specifically targets Arabic, Bulgarian, German, Italian and English by fine-tuning a BERT model adapted for SA, achieving competitive results in subjectivity classification across these languages.

SD in specific languages has also been explored in isolation. Arabic-focused research includes *ThatiAR* [16], an Arabic dataset for SD composed of approximately 3.600 sentences. In this study, the authors benchmarks fine-tuning and prompting techniques with different LLMs, showing that Arabic-adapted pre-tained language models are effective for SD in Arabic and outperform multilingual LLMs. Several teams from past CLEF's CheckThat! lab editions have developed monolingual Transformer-based systems for German and Italian that achieved competitive results. For instance, at the 2023 edition, the DWReCO team [17] fine-tuned BERT-based models such as German BERT and BERTurk, demonstrating that different subjective styles are effective across languages. Additionally, at the 2024 edition, the HYBRINFOX team [18] used a hybrid system combining a fine-tuned RoBERTa model, a frozen sentence BERT (sBERT) and several scores calculated by the English version of the VAGO [19] expert system, to enhance SD in Italian. This system outperformed multilingual baselines in that track.

Claim Extraction, which is understood as the detection and extraction of claims present in a text with the aim of making them suitable for verification, is a main task in automatic fact-checking systems, especially in environments such as social networks, where claims often appear in informal, implicit or fragmented language. Early approaches focused on identifying check-worthy sentences using supervised classifiers trained with linguistic features such as the presence of digits, entities, modal expressions or lexical n-grams in the text. Later studies [20, 21] also incorporated discourse analysis features or stylistic properties, and explored unsupervised approaches to identify check-worthy sentences present in a text. With the rise of deep learning, transformer-based pre-trained language models have recently dominated the field, both in performance and focus of attention. However, these approaches are limited, as they rely on whole-sentence extraction, which does not guarantee that statements are clear or self-contained, as extracting whole sentences as statements may include irrelevant or context-dependent content [22].

Given that many traditional methods were limited to extracting the original sentence, the extracted statements were usually embedded in noise, conversational context or colloquial language, due to these limitations the notion of Claim Normalization (ClaimNorm) [23] emerged, which is understood as the

process that aims to extract (extraction) and reformulate the statements present in a text in a clean, self-contained and verifiable way (normalization), eliminating ambiguities, contextual references or rhetorical ornaments. To perform this normalization, the approaches employed are based on the use of generative models based on encoder-decoder architectures, models such as T5 and Flan-T5, which, given an input text sequence, are capable of producing an output text sequence. To solve this task, these models are used to perform an abstractive summarization of the text content, extracting the statements present in the text and generating their normalized form [22, 23, 24].

3. System Overview

Figure 1 shows the overall system architecture for the Subjectivity subtask, where encoder-only Transformer models are used as the basis for performing fine-tuning for a text classification task. In this task, the whole input text is labeled with a binary value: 1 if the text is subjective, that is, if it contains opinions, feelings, or personal appreciations, and 0 otherwise.

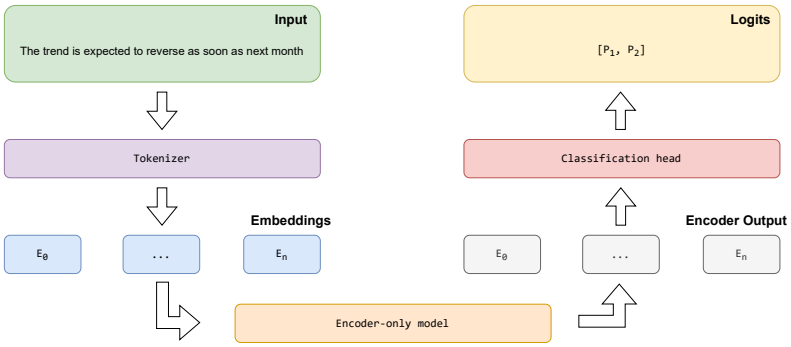

Figure 1: System architecture for the Subjectivity subtask.

In the monolingual settings, we fine-tuned language-specific encoder-only models to better capture the linguistic nuances of each language, thereby ensuring more precise and specialized performance. Therefore, for each language a model pre-trained on data from that specific linguistic domain was selected. This strategy ensures that each model is aligned with the linguistic characteristics and textual structures of its respective language, which contributes to improved performance in language-specific evaluation settings.

Specifically, for each language we selected pre-trained models tailored to the characteristics of each language. For Arabic, we used MARBERTv2 [25], a model trained on Arabic social media text to better represent dialectal and informal language. This model outperforms both multilingual models and AraBERT, while remaining more energy efficient across diverse downstream tasks on the ARLUE dataset benchmark. For Bulgarian, we used BERTić [26], which was trained on large-scale Bosnian, Croatian, Montenegrin and Serbian corpora with support for South Slavic languages. It achieved superior performance on tasks such as POS tagging, NER, geolocation and commonsense reasoning when compared to multilingual models. For German, we used German BERT [27], which was trained on large German corpora and outperformed both multilingual models and earlier German BERT variants on NER, text classification and hate speech detection tasks. Most notably, it performed similarly to GottBERT, which set new benchmarks on GermEval and CoNLL 2003. For Italian, we used BERTino [28], a DistilBERT model variant that was trained on a large, general-domain Italian corpus. BERTino achieved F1 scores comparable to those of a full-sized BERT model across multiple tasks, including POS tagging, NER and sentence classification, while reducing fine-tuning and inference time by half. Finally, for English, we fine-tuned the widely adopted RoBERTa-Base [29] model, which was pre-trained on a large English corpus. Fine-tuning was carried out using only the training data corresponding to each language, allowing the models to specialize in language-specific features such as syntax, morphology and

lexical semantics. This specialization enables more accurate and contextually appropriate predictions within each language.

For multilingual and zero-shot scenarios, we employed XLM-RoBERTa-Large [30], a Transformer-based multilingual encoder model trained on 2.5 TB of filtered CommonCrawl data covering 100 languages. This model was pre-trained using the Masked Language Modeling (MLM) objective, where 15% of the tokens in the input are randomly masked and predicted from their surrounding context. Unlike recurrent architectures such as RNNs, or autoregressive models like GPT that mask future tokens, XLM-RoBERTa processes input bidirectionally. This enables the model to build contextualized representations that are well-suited for multilingual tasks, and particularly effective in zero-shot cross-lingual transfer settings. The fine-tuning was performed using the concatenated training data from all available languages, allowing the model not only to learn generalizable patterns across linguistic boundaries, but also to acquire language-specific patterns tied to the multilingual setting, based on the languages seen during training. This dual capability enhances the model's robustness in zero-shot settings while preserving its effectiveness in multilingual contexts.

We used the HuggingFace Transformers library to perform the fine-tuning for the Subjectivity subtask, formulating it as a binary text classification problem. Each input text was assigned a single label: 1 for subjective texts, and 0 for objective ones. For each setting (monolingual, multilingual and zero-shot) we loaded the appropriate pre-trained encoder-only model using AutoModelForSequenceClassification, together with its corresponding tokenizer. Thus, the classification head added on top of the encoder consists of a single linear layer on top of the [CLS] token, which produces logits for the two classes.

Fine-tuning was performed using the Trainer API, with cross-entropy loss as the optimization objective. Standard preprocessing steps included tokenizing the input texts, truncating or padding them to a fixed maximum length, and batching them efficiently for GPU training. Special tokens and padding were handled automatically by the tokenizer, and no additional label masking was required since classification is performed at the sequence level.

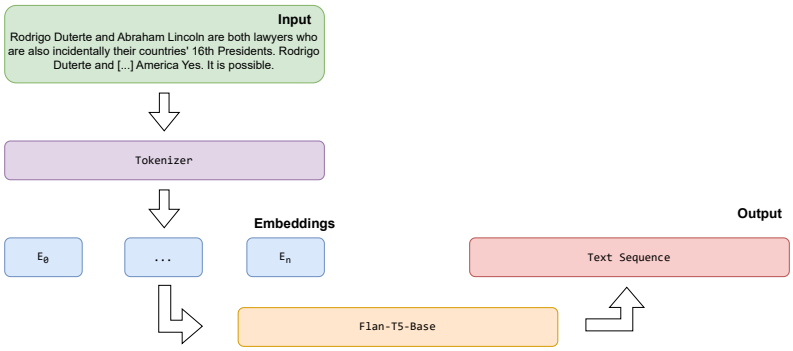

Figure 2: System architecture for the Claim Extraction & Normalization subtask.

Figure 2 shows the overall system architecture for the Claim Extraction & Normalization subtask, where we fine-tune Flan-T5-Base [10], a multilingual Text-to-Text Transfer Transformer model, to generate normalized claims from social media posts. In this sequence-to-sequence setup, the model receives as input a text that potentially contains a claim, and learns to produce as output a normalized version of that claim, that is, a clear, concise and generalized reformulation suitable for downstream fact-checking.

In the monolingual settings, we fine-tuned separate instances of the Flan-T5-Base model, one for each language, to better capture the linguistic nuances and stylistic conventions specific to that language. We reused the same multilingual architecture and performed language-specific fine-tuning using only the training data available for that particular language. This approach allows each Flan-T5-Base instance to specialize in the syntax, morphology and vocabulary of a single language, improving its ability to generate accurate and well-formed normalized claims. By isolating the training per language, the models are encouraged to internalize language-specific patterns, leading to more precise and contextually

appropriate outputs in language-specific evaluation settings.

For zero-shot scenarios a instance of the Flan-T5-Base model was fine-tuned using the concatenated training data from all available languages. This multilingual training strategy enabled the model to learn both generalizable patterns that transfer across linguistic boundaries and language-specific features grounded in the languages seen during training. As a result, the model gains a dual capability: it becomes robust to cross-lingual variation while retaining sensitivity to individual language characteristics, a key advantage for zero-shot generalization.

We used the HuggingFace Transformers library to fine-tune the model for the Claim Extraction & Normalization subtask, formulating it as a sequence-to-sequence generation task. Given an input text potentially containing a claim, the model was trained to generate a normalized version of that claim. For each setting (monolingual and zero-shot) we loaded the model using AutoModelForSeq2SeqLM, along with its corresponding tokenizer. The model architecture naturally supports text-to-text generation, with no need for a separate classification head.

Fine-tuning was performed using the Seq2SeqTrainer API, optimizing with cross-entropy loss between the generated output tokens and the gold normalized claims. Preprocessing included tokenizing both inputs and targets, padding or truncating them to fixed maximum lengths, and dynamically batching them for efficient training on GPU. The tokenizer handled special tokens and padding automatically, and label masking was applied where appropriate to ignore padding tokens during loss computation.

4. Experimental Setup

For Subjectivity subtask, we based our experimentation solely on the training and development sets provided by the task organizers. Each language came with three annotated subsets: train, dev and dev_test, whose number of examples can be seen in Table 1.

Table 1
Subjectivity subtask dataset distribution

Language	Train set	Development set	Development test set
Arabic	2,446	742	748
Bulgarian	729	467	250
German	800	491	337
Italian	1,613	667	513
English	830	462	484

In the monolingual settings, we fine-tuned distinct pre-trained language models for each language using only the training data corresponding to that language. The fine-tuning process was performed for 5 epochs with a learning rate of `5e-6`, a weight decay of 0.01 and a batch size of 16. For the multilingual and zero-shot scenarios, a single instance of the XLM-RoBERTa-Large model was fine-tuned on the concatenation of all available training data across languages, using the same hyperparameter configuration. This consistent setup ensured a fair comparison across experimental settings.

For Claim Extraction & Normalization subtask, we have based our experimentation solely on the training and development sets provided by the task organizers. Each language came with two annotated subsets: train and dev, whose number of examples can be seen in Table 2.

In the monolingual settings, we fine-tuned distinct instances of the Flan-T5-Base model for each language using only the training data corresponding to that language. The fine-tuning process was performed for 15 epochs with a learning rate of `5e-7`, a weight decay of 0.01 and a batch size of 2. During training and evaluation, we enabled sequence generation by setting `predict_with_generate = True` in the Seq2SeqTrainer configuration. Additionally, the maximum length for generated sequences was set to 128 tokens (`generation_max_length = 128`) as required in the task. For the zero-shot scenarios, a single instance of the Flan-T5-Base model was fine-tuned on the concatenation of all

Table 2
Claim Extraction & Normalization subtask dataset distribution

Language	Train set	Development set
Arabic	470	118
Indonesian	540	137
German	386	101
French	1174	147
Polish	163	41
Thai	244	61
Tamil	102	50
Punjabi	445	50
Marathi	137	50
Hindi	1081	50
Portugese	1735	223
Spanish	3458	439
English	11374	1171

available training data across languages, using the same hyperparameter configuration. This consistent setup ensured a fair comparison across experimental settings.

5. Results

Table 3 presents the official results obtained in the Subjectivity subtask, reported by language and evaluation context (monolingual, multilingual and zero-shot). Our approach achieved competitive performance across all settings, consistently outperforming the baseline in every case.

Table 3
Subjectivity subtask results reported by language and evaluation context

Language	Winner team	F1-score of the winner team	Our place in the ranking	Our F1-score	Baseline
Monolingual context					
Arabic	CEA-LIST	0.6884	2nd	0.5903	0.5133
German	smollab	0.8520	10th	0.7324	0.6960
Italian	XplaiNLP	0.8104	4th	0.7703	0.6941
English	msmadi	0.8052	4th	0.7604	0.5370
Multilingual context					
Multilingual	TIFIN India	0.7550	7th	0.7074	0.6390
Zero-Shot context					
Polish	CEA-LIST	0.6922	7th	0.5763	0.5719
Ukrainian	CSECU-Learners	0.6424	6th	0.6210	0.6296
Romanian	msmadi	0.8126	5th	0.7793	0.6461
Greek	AI Wizards	0.5067	4th	0.4831	0.4159

In the monolingual setting, our fine-tuned models ranked among the top positions for most languages. Notably, in Arabic, we obtained an F1-score of 0.5903 and ranked 2nd, only behind the winning team (CEA-LIST). In Italian and English, we also achieved strong results, placing 4th in both languages with F1-scores of 0.7703 and 0.7604, respectively. These results highlight the benefit of leveraging language-specific pre-trained models, which are better equipped to handle nuances such as morphology, syntax and stylistic variation. While the result for German (10th place) shows room for improvement, our system still exceeded the baseline by a margin of 0.0364 points.

In the multilingual setting, where all languages were combined during training, our model achieved an F1-score of 0.7074 and ranked 7th, outperforming the multilingual baseline and confirming that cross-lingual transfer is viable when training jointly across languages. However, results were generally lower than in the best-performing monolingual configurations, reinforcing the advantage of language-specific fine-tuning when sufficient training data is available.

In the more challenging setting, the zero-shot one, where no training data from the target language was used, our model still surpassed the baseline in all cases. Particularly strong results were observed for Romanian (F1-score 0.7793, 5th place) and Greek (F1-score 0.4831, 4th place). While results in Polish and Ukrainian were closer to the baseline, these results suggest that the multilingual fine-tuned model was able to generalize effectively to unseen languages, especially when they share features with those observed during training.

In summary, our results show that language-specific fine-tuning leads to better performance in monolingual scenarios, while multilingual fine-tuning enables effective cross-lingual generalization, especially in zero-shot conditions. This confirms the complementary strengths of both strategies depending on the availability of annotated resources for each language.

Table 4 presents the official results for the Claim Extraction & Normalization subtask, reported by language and evaluation context (monolingual and zero-shot). While our system did not achieve top positions in the final rankings, it consistently produced valid outputs across all languages, showing the feasibility of using a unified, multilingual generative model for this challenging task.

Table 4
Claim Extraction & Normalization subtask results reported by language and evaluation context (monolingual, multilingual and zero-shot)

Language	Winner	F1-score of the winner	Our place in the ranking	Our F1-score
Monolingual context				
Arabic	tatiana.anikina	0.5037	8th	0.0003
Indonesian	DSGT-CheckThat	0.5650	7th	0.1305
German	DSGT-CheckThat	0.3859	9th	0.1039
French	DSGT-CheckThat	0.5273	10th	0.1649
Polish	DSGT-CheckThat	0.4065	9th	0.0742
Thai	DSGT-CheckThat	0.5859	7th	0.014
Tamil	tatiana.anikina	0.6316	9th	0.0043
Punjabi	tatiana.anikina	0.3307	9th	0.0097
Marathi	tatiana.anikina	0.3888	9th	0.0877
Hindi	tatiana.anikina	0.3275	11th	0.0132
Portugese	DSGT-CheckThat	0.5770	7th	0.1898
Spanish	DSGT-CheckThat	0.6077	9th	0.2048
English	tatiana.anikina	0.4569	16th	0.1660
Zero-Shot context				
Dutch	tatiana.anikina	0.2001	6th	0.0817
Romanian	tatiana.anikina	0.2950	6th	0.0779
Bengali	tatiana.anikina	0.3777	8th	0.0451
Telugu	tatiana.anikina	0.5257	9th	0.0269
Korean	tatiana.anikina	0.1339	6th	0.0014
Greek	tatiana.anikina	0.2619	6th	0.0062
Czech	tatiana.anikina	0.2519	6th	0.0544

In the monolingual setting, we fine-tuned individual instances of Flan-T5 for each language, obtaining modest F1-scores compared to those of the best performing systems. The best results were observed in Spanish (F1 = 0.2048), Portuguese (F1 = 0.1898) and French (F1 = 0.1649), suggesting that the model adapts better to languages with more lexical similarity to English or with better pretraining coverage. These results also reflect the difficulty of the task, which requires not only claim detection but also accurate reformulation in a normalized format.

In the zero-shot setting, where no language-specific fine-tuning was performed, the model demonstrated low generalization ability to unfamiliar languages. For example, in Dutch, Romanian and Czech, the system achieved F1-scores above 0.05, indicating that multilingual training allows a very low, almost non-existent degree of cross-lingual transfer. Performance remains limited overall, highlighting the complexity of generating accurate and semantically faithful normalized claims in low-resource or unseen languages.

Although the system ranked towards the lower end in most cases, these results provide a valuable baseline for future work, underscoring the need for a more sophisticated approach. In summary, our experiments confirm the importance of language-specific adaptation and the potential of multilingual generative models as a foundation for claim normalization in diverse linguistic contexts.

6. Conclusions and Further Work

In this work, we addressed two complementary subtasks: Subjectivity Classification and Claim Extraction & Normalization, focusing on both monolingual, zero-shot and multilingual scenarios. For the Subjectivity subtask, we adopted an encoder-only fine-tuning strategy, using language-specific models in the monolingual setting and XLM-RoBERTa for multilingual and zero-shot settings. The results demonstrated that language-specific fine-tuning can effectively capture linguistic nuances, leading to improved performance when annotated data is available. In contrast, multilingual fine-tuning offered greater robustness in both multilingual and zero-shot scenarios, allowing the model to generalize to unseen languages, even though with a slight decrease in accuracy.

For the Claim Extraction & Normalization subtask, we proposed a generative sequence-to-sequence approach based on Flan-T5. We fine-tuned separate instances of the model per language for monolingual experiments, and a single multilingual model for zero-shot evaluation. While our system did not achieve top rankings, it consistently produced valid outputs across all languages, including low-resource ones. These results confirm the viability of generative multilingual models for structured text transformation tasks, despite the inherent difficulty of aligning informal, noisy inputs with formalized claim outputs.

Several directions remain open for future work. In the Subjectivity subtask, exploring data augmentation strategies or using additional labeled corpora may also enhance robustness, specifically in low-resource languages. In the Claim Extraction & Normalization subtask, future efforts should focus on improving the semantic fidelity and coherence of generated claims. This could be achieved through techniques such as contrastive learning, reinforcement learning with human feedback, or explicit semantic similarity constraints. Additionally, instruction tuning or prompt-based methods may help improve zero-shot performance without the need for separate fine-tuning per language.

Acknowledgments

This work is part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MCIN/AEI/10.13039/501100011033 and the European Fund for Regional Development (ERDF)-a way to make Europe. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL for grammatical and spelling correction. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] N. Firdaus, J. Jumroni, A. Aziz, E. Sumartono, A. Purwanti, The Influence of Social Media, Misinformation, and Digital Communication Strategies on Public Perception and Trust, The Journal of Academic Science 1 (2024) 131–138.

[2] N. Hassan, B. Adair, J. Hamilton, C. Li, M. Tremayne, J. Yang, C. Yu, The Quest to Automate Fact-Checking, Proceedings of the 2015 Computation + Journalism Symposium (2015).

[3] V. Setty, Surprising Efficacy of Fine-Tuned Transformers for Fact-Checking over Larger Language Models, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 2842–2846.
[4] Z. Guo, M. Schlichtkrull, A. Vlachos, A Survey on Automated Fact-Checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.
[5] 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.
[6] 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: [31], 2025, pp. –.
[7] M. Sundriyal, T. Chakraborty, P. Nakov, Overview of the CLEF-2025 CheckThat! Lab Task 2 on Claim Normalization, in: [31], 2025, pp. –.
[8] V. Venktesh, V. Setty, A. Anand, M. Hasanain, B. Bendou, H. Bouamor, F. Alam, G. Iturra-Bocaz, P. Galuščáková, Overview of the CLEF-2025 CheckThat! Lab Task 3 on Fact-Checking Numerical Claims, in: [31], 2025, pp. –.
[9] S. Hafid, Y. S. Kartal, S. Schellhammer, K. Boland, D. Dimitrov, S. Bringay, K. Todorov, S. Dietze, Overview of the CLEF-2025 CheckThat! Lab Task 4 on Scientific Web Discourse, in: [31], 2025, pp. –.
[10] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling Instruction-Finetuned Language Models, Journal of Machine Learning Research 25 (2024) 1–53.
[11] A. Rodríguez, E. Golobardes, J. Suau, Tonirodriguez at CheckThat! 2024: Is it Possible to Use Zero-Shot Cross-Lingual Methods for Subjectivity Detection in Low-Resources Languages?, in: CEUR Workshop Proceedings, volume 3740, CEUR-WS, 2024, pp. 590–597.
[12] B. Pang, L. Lee, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004, pp. 271–es.
[13] R. Satapathy, S. R. Pardeshi, E. Cambria, Polarity and Subjectivity Detection with Multitask Learning and BERT Embedding, Future Internet 14 (2022) 191.
[14] C. Banea, R. Mihalcea, J. Wiebe, Multilingual Subjectivity: Are More Languages Better?, in: Proceedings of the 23rd international conference on computational linguistics (Coling 2010), 2010, pp. 28–36.
[15] M. R. Biswas, A. T. Abir, W. Zaghouani, Nullpointer at CheckThat! 2024: Identifying Subjectivity from Multilingual Text Sequence, in: CEUR Workshop Proceedings, volume 3740, CEUR-WS, 2024, pp. 361–368.
[16] R. Suwaileh, M. Hasanain, F. Hubail, W. Zaghouani, F. Alam, ThatiAR: Subjectivity Detection in Arabic News Sentences, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 19, 2025, pp. 2587–2602.
[17] I. B. Schlicht, L. Khellaf, D. Altiok, Dwreco at CheckThat! 2023: Enhancing Subjectivity Detection through Style-based Data Sampling, arXiv preprint arXiv:2307.03550 (2023).
[18] M. Casanova, J. Chanson, B. Icard, G. Faye, G. Gadek, G. Gravier, P. Égré, HYBRINFOX at CheckThat! 2024-Task 2: Enriching BERT Models with the Expert System VAGO for Subjectivity Detection, in: CLEF 2024-Conference and Labs of the Evaluation Forum, 2024, pp. 1–9.
[19] B. Icard, G. Atemezing, P. Égré, VAGO: un outil en ligne de mesure du vague et de la subjectivité, in: Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (PFIA 2022), 2022, pp. 68–71.
[20] X. Zhou, B. Wu, P. Fung, Fight for 4230 at CheckThat! 2021: Domain-Specific Preprocessing and Pretrained Model for Ranking Claims by Check-Worthiness, in: CLEF (Working Notes), 2021, pp. 681–692.

[21] 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), 2011, pp. 48–57.

[22] Z. Deng, M. Schlichtkrull, A. Vlachos, Document-level Claim Extraction and Decontextualisation for Fact-Checking, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 11943–11954.

[23] 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, 2023, pp. 6594–6609.

[24] H. Ullrich, T. Mlynář, J. Drchal, Claim Extraction for Fact-Checking: Data, Models, and Automated Metrics, arXiv preprint arXiv:2502.04955 (2025).

[25] M. Abdul-Mageed, A. Elmadany, E. M. B. Nagoudi, ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 7088–7105. URL: https://aclanthology.org/2021.acl-long.551. doi:10.18653/v1/2021.acl-long.551.

[26] N. Ljubešić, D. Lauc, BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian, in: Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, Association for Computational Linguistics, Kiyv, Ukraine, 2021, pp. 37–42. URL: https://www.aclweb.org/anthology/2021.bsnlp-1.5.

[27] B. Chan, S. Schweter, T. Möller, German's Next Language Model, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6788–6796.

[28] M. Muffo, E. Bertino, BERTino: an Italian DistilBERT model, Computational Linguistics CLiC-it 2020 (2020) 317.

[29] 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, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[30] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[31] 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.