Entity-aware Cross-lingual Claim Detection for Automated Fact-checking

Rrubaa Panchendrarajan, Arkaitz Zubiaga
School of Electronic Engineering and Computer Science
Queen Mary University of London
{r.panchendrarajan, a.zubiaga}@qmul.ac.uk

Abstract

Identifying claims requiring verification is a critical task in automated fact-checking, especially given the proliferation of misinformation on social media platforms. Despite notable progress, challenges remain—particularly in handling multilingual data prevalent in online discourse. Recent efforts have focused on fine-tuning pre-trained multilingual language models to address this. While these models can handle multiple languages, their ability to effectively transfer cross-lingual knowledge for detecting claims spreading on social media remains under-explored. In this paper, we introduce *EX-Claim*, an entity-aware cross-lingual claim detection model that generalizes well to handle multilingual claims. The model leverages entity information derived from named entity recognition and entity linking techniques to improve the language-level performance of both seen and unseen languages during training. Extensive experiments conducted on three datasets from different social media platforms demonstrate that our proposed model stands out as an effective solution, demonstrating consistent performance gains across 27 languages and robust knowledge transfer between languages seen and unseen during training.

1 Introduction

Automated fact-checking is an emerging research task dedicated to combating misinformation, particularly prevalent on social media, and comprises several key stages: claim detection, claim prioritization, evidence retrieval, and claim validation (Zeng et al., 2021). While research in the task is progressing rapidly, identifying and validating claims related to global concerns requires a fact-checking pipeline capable of processing claims written in multiple languages. However, developing multilingual solutions for fact-checking research is not straightforward due to the availability of limited training data, especially for low-resource languages (Panchendrarajan and Zubiaga, 2024). This necessitates advancements in multilingual fact-checking research to leverage limited multilingual training data for developing generalized solutions that can effectively transfer knowledge between languages.

This research focuses on the first component of the fact-checking pipeline: verifiable claim detection. A *verifiable* claim is defined as a statement expressing facts, excluding personal experience and private knowledge (Panchendrarajan and Zubiaga, 2024). Existing works on verifiable claims detection are predominantly focused on monolingual solutions (Prabhakar et al., 2020; Suri and Dudeja, 2022; Henia et al., 2021; Hussein et al., 2021), with limited attention given to addressing the multilingual nature of the problem. Especially, fine-tuning multilingual language models such as mBERT (Alam et al., 2021b; Uyangodage et al., 2021; Panda and Levitan, 2021) and XLM-R (Alam et al., 2021b; Hüsünbeyi et al., 2022; Savchev, 2022; Alam et al., 2021a) is observed as a common solution. While these models can handle multiple languages, none of them adequately validate the cross-lingual knowledge-transferring capability of these models, particularly in identifying claims written in languages that were not seen during training. To address this challenge, we propose *EX-Claim*, a cross-lingual claim detection model that can effectively identify verifiable claims regardless of their language. Our model broadens the ability to perform across a diverse set of languages beyond those seen in training by incorporating entity-centric information for knowledge transfer. Our approach leverages the limited multilingual data available in claim detection research to develop a robust cross-lingual model that generalizes beyond the languages seen in the training phase.

We build our model on the assumption that factual claims circulating in social media often revolve around entities. In particular, entity types and their statuses such as their popularity may affect the

Findings of the Association for Computational Linguistics: EACL 2026, pages 17–33
March 24-29, 2026 ©2026 Association for Computational Linguistics

verifiability of a claim. For instance, consider the following three statements.

- *S1: **X** visited **China** to attend **EMNLP***
- *S2: **Keir Starmer** visited **France** to attend **Euro Con***
- *S3: **Keir Starmer** visited his home in the **UK***

Statement $S1$ can be deemed unverifiable from the perspective of a fact-checker, as it describes a personal experience of an unknown individual *X*. However, a similar context of attending an event expressed in $S2$ is verifiable as the popular entity *Keir Starmer* (the UK Prime Minister at the time of writing) attending an official event *Euro Con* becomes verifiable. However, changing the popular entities *Keir Starmer* or *Euro Con* with another popular *person* or *event* entity will not change the verifiability status of this claim. At the same time, changing the *event* with a private location such as *home* reduces the verifiability of the claim $S3$ even if a popular entity is involved. This motivates our intuition that entities, their types, their popularity, and the relationship between them affect the verifiability status of the claims.

Our proposed *EX-Claim* builds on this intuition by leveraging entity information to effectively identify verifiable claims in a cross-lingual environment. Our model incorporates entity information such as type and popularity derived from named entity recognition (NER) and entity linking (EL). Extensive experiments on three datasets from different social media platforms comprising 27 languages show that *EX-Claim* stands out as the effective model with consistent performance gain across multiple languages and robust knowledge transfer between languages seen and unseen in training. We make the following key contributions:

- We propose *EX-Claim*[1], a cross-lingual claim detection model for identifying verifiable claims from social media which can effectively transfer knowledge across languages seen and unseen during training.
- We enhance language-level performance by leveraging entity information in claims through NER and EL.
- We conduct extensive experiments to evaluate *EX-Claim* across three datasets representing different social media platforms and encompassing 27 languages, including synthetic data created via machine translation.

[1]The source code and synthetic dataset are available at `https://github.com/RubaP/Ex-Claim`

2 Related Work

Identifying verifiable claims and check-worthy claims are often conflated in the literature and collectively referred to as claim detection. While check-worthy claim detection is inherently subjective, depending on additional factors such as popularity, impact, or timeliness, both tasks are often addressed using the same or closely related approaches (Panchendrarajan and Zubiaga, 2024).

A common approach to identifying multilingual claims shared on social media involves fine-tuning pre-trained multilingual language models such as mBERT (Alam et al., 2021b; Uyangodage et al., 2021; Panda and Levitan, 2021; Zengin et al., 2021; Hasanain and Elsayed, 2022) and XLM-R (Alam et al., 2021b,a) and developing language-specific models to achieve optimal performance Hüsünbeyi et al. (2022); Savchev (2022); Eyuboglu et al. (2023). Most of these works (Alam et al., 2021b,a; Uyangodage et al., 2021; Baris-Schlicht et al., 2021; Du et al., 2022) fine-tune the language model using combined training data of multiple languages and evaluate the capability of the model in identifying claims written in the same set of languages. Very few studies (Panda and Levitan, 2021; Zengin et al., 2021; Hasanain and Elsayed, 2022) explore the model's ability to transfer knowledge across language pairs, and these studies are limited to a narrow selection of languages.

With the recent attention to large language models (LLM), Agrestia et al. (2022) utilized GPT-3 to fine-tune the model in each language to develop a language-specific claim detection model. Interestingly, the results revealed that transformer-based architectures like BERT remain competitive with, and in some cases outperform the large language models (LLMs). Following these findings, further research (Sawiński et al., 2023; Lewoniewski et al., 2024; Li et al., 2024) explored leveraging LLMs for claim detection by modeling the task as a text generation problem. While these models demonstrate superior performance in the English language, their ability to handle multilingual claims for the claim detection task has yet to be explored.

Beyond directly using multilingual language models, various methods have been applied to tackle limited training data and noisy social media text. This includes the extraction of platform-specific textual features (Alam et al., 2021b), data augmentation techniques such as machine translation (Zengin et al., 2021; Savchev, 2022) and

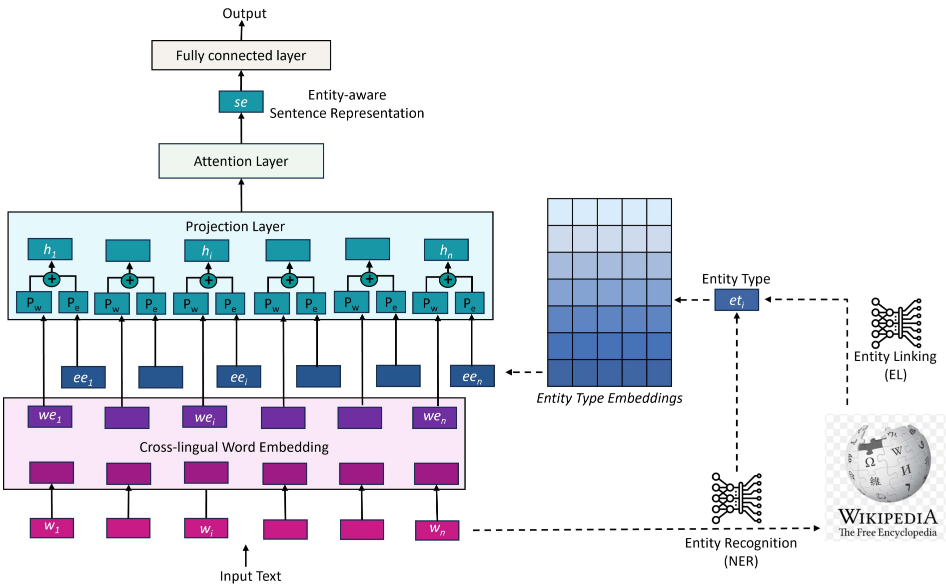


Figure 1: *EX-Claim :* Entity-aware Cross-lingual Claim Detection

sampling (Zengin et al., 2021) and multitasking (Baris-Schlicht et al., 2021; Du et al., 2022). While these techniques are explored in multilingual settings, none of them have been shown to improve cross-lingual knowledge-transferring capabilities, which our study performs comprehensively for the first time looking at 27 different languages through the exploration of entity-based knowledge transfer.

3 Methodology

Given a statement written in any language, we classify it as verifiable or non-verifiable. Figure 1 illustrates the proposed solution, *EX-Claim*: entity-aware cross-lingual claim detection framework, whose components are discussed next.

3.1 Cross-lingual Word Embedding

We use XLM-R (Kalyan et al., 2021) to extract the cross-lingual vector representation of the words. XLM-R was trained on CommonCrawl data supporting 100 languages, and generates embeddings of size 768 for each tokenized word.

3.2 Named Entity Recognition (NER)

Identifying entities mentioned in the text is essential to extract further information such as its type and popularity. We use state-of-the-art multilingual NER model MultiNERD (Tedeschi and Navigli, 2022) for identifying named entities. As the model wasn't released by the authors, we fine-tuned the MultiNERD model with recommended parameters using their source code and training data. MultiNERD recognizes 15 types of fine-grained named entities. Refer to Appendix A.1 for the named entity types and NER statistics.

3.3 Entity Linking (EL)

We consider an entity *popular* if it can be linked to a Wikipedia article. Hence, we run EL for all entities identified in the previous stage to determine if an entity has an associated Wikipedia article that can be linked to. To do this, we make use of a state-of-the-art multilingual EL model, mGENERE (De Cao et al., 2022) for linking entities. Given an entity-tagged sentence, mGENERE outputs a ranked list of n Wikipedia page titles along with their associated log probabilities, from which we make use of the top item with the highest probability to determine if indeed the entity has a relevant Wikipedia article that can be linked to. If the log probability is above a threshold, we consider that entity as popular.

During parameter tuning, we observed that the variation in threshold did not significantly affect the performance of the claim detection models that utilize EL. This is likely because most named entities were already recognized as Wikipedia entities, making the additional information redundant—possibly due to the NER model's training data being partially sourced from Wikipedia. Consequently, we set the EL threshold to -0.15 (probability 0.861), which notably reduced the number of linked entities (refer to Figure 5 in Appendix A.2), ensuring that only highly probable entity links are used to determine the popularity. Refer to Appendix A.2 for the EL statistics.

3.4 Entity Type Embedding

Similar to the word embedding, an entity type can be represented in a dense vector representation, enabling the model to capture entity type information required for the task as a dense vector. This is achieved by adding an embedding layer to the claim detection model, which converts each entity type to an entity type vector. Based on the entity type leveraged, we experiment with the following two variants of *EX-Claim*:

- *EXN-Claim*: Uses 15 named entity types recognized by the NER model, along with the *other* category for non-entities—resulting in a total of 16 entity types.
- *EXP-Claim*: Uses both NER and the popularity of an entity. Each named entity can be either popular or unpopular, doubling the size of the entity types. This results in 31 entity types (2*15 + 1).

As an ablated variant, we experiment with *X-Claim* category which does not use any entity-type information to show the effectiveness of utilizing entity-type information for verifiable claim detection. Except for the *X-Claim* category, each word in the input sequence is identified with an entity type index et_i according to one of the entity type categories. This information is fed into the entity type embedding layer to obtain a dense vector representation ee_i for each word in the input sequence.

3.5 Entity-aware Claim Detection

Given an input word sequence $(w_1, w_2, \ldots, w_n)$, we extract the cross-lingual word embeddings $(we_1, we_2, \ldots, we_n)$, and the entity type index $(et_1, et_2, \ldots, et_n)$. Here, $et_i \in \mathbb{R}^k$ represents the entity type of a word as a one-hot vector, where k indicates the number of entity types. These entity-type indices are used to extract entity-type embeddings as follows,

$$EE(et_1, et_2, ...et_n) = (ee_1, ee_2, ...ee_n) \tag{1}$$

Once the entity-type embeddings are obtained, we project both word embeddings and entity-type embeddings into a common vector space and add the projected vectors at the word level as follows,

$$h_i = P_w \cdot we_i + P_e \cdot ee_i \in \mathbb{R}^{d_p} \tag{2}$$

This projection layer is responsible for converting the input features into task-specific, low-dimensional vectors and incorporating entity-type information into the word embedding.

Notation	Description
w_i	i^{th} word from the input sequence
n	Number of words in the input sequence
$we_i \in \mathbb{R}^{d_w}$	Word embedding of i^{th} word
d_w	Dimension of word embedding
$et_i \in \mathbb{R}^k$	One-hot vector indicating entity type of i^{th} word
k	Number of entity types
$ee_j \in \mathbb{R}^{d_e}$	Entity type embedding of j^{th} entity type
d_e	Dimension of entity type embedding
d_p	Projection layer dimension
$P_w \in \mathbb{R}^{d_w \times d_p}$	Word embedding projection matrix
$P_e \in \mathbb{R}^{d_e \times d_p}$	Entity type embedding projection matrix
$h_i \in \mathbb{R}^{d_p}$	Projected input representation
H	Projected input sequence $(h_1, h_2, ...h_n)$
$se \in \mathbb{R}^{d_p}$	Sentence embedding
C	Number of class labels
$y \in \mathbb{R}^C$	Probability distribution predicted by the model for the input sequence
$\hat{y} \in \mathbb{R}^C$	Ground truth probability distribution of the input sequence
$EE \in \mathbb{R}^{k \times d_e}$	Entity embedding layer
$Attn$	Attention layer

Table 1: List of Key Notations

To attend to the important words indicating the verifiability of claims, we adopt a simplified variant of self-attention (Vaswani et al., 2017). Specifically, the projected hidden states $H = (h_1, h_2, ...h_n)$ are used directly as query, key, and value representations, eliminating separate Q–K–V projections and thereby reducing the number of parameters. Empirically, we find that this formulation sufficiently captures the dependencies needed for claim detection, while keeping the model lightweight. The attention over the hidden states is computed as:

$$Attn(H) = softmax(\frac{HH^T}{\sqrt{d_p}}) \cdot H \tag{3}$$

The attention layer $Attn$ generates attended context vectors for each element in the input sequence. The mean value of this context vector is obtained as the sentence representation se as follows,

$$se = \frac{1}{n} \sum Attn(H) \in \mathbb{R}^{2d_p} \tag{4}$$

Finally, the sentence embedding se is fed into a fully connected layer followed by a softmax function to generate a probability distribution y over the output classes as follows,

$$out = W_o se + b_o \in \mathbb{R}^C \tag{5}$$

$$y_j = \frac{exp(out_j)}{\sum_j^C exp(out_j)} \in \mathbb{R}^C \tag{6}$$

Given the true label distribution $\hat{y} \in \mathbb{R}^C$, the entity type embedding layer, projection layer, and attention layer are trained by optimizing the cross-entropy loss.

4 Experiment Setting

4.1 Dataset

We use the CheckThat! 2022 (Nakov et al., 2022) Task 1B dataset for training and testing the verifiable claim detection models. This dataset contains tweets related to COVID-19 in 5 languages, labeled as verifiable or unverifiable. Further, it contains four partitions: Train, Dev, Dev-test, and Test. We used the Train partition of all languages except Arabic for training and used the Dev partition as validation data. The Arabic data was excluded during the training as the initial experiments yielded poor performance in all languages including Arabic. However, Arabic was retained in the test partition to enable a fair and comprehensive evaluation. Further, the Dev-Test partition was used to find the optimal parameters.

To evaluate the models' ability to identify claims written in a broad range of unseen languages, we generated synthetic data from the CheckThat! Test partition for 18 unseen languages. For each, we randomly sampled 250 instances and translated them using Microsoft Azure AI translator (Junczys-Dowmunt, 2019), limiting samples to match the test dataset size. Given the scarcity of multilingual machine translation benchmarks for social media, translation quality was assessed on the FLORES-200 benchmark (Team et al., 2022), with a CHRF score of 54.35. Further details are in Appendix B.2.

To evaluate the generalization capability of the proposed model in identifying claims discussing unseen topics as well as unseen language, we report the performance in Kazemi 2021 (Kazemi et al., 2021). This dataset contains WhatsApp tipline and social group messages written in 5 languages. Refer to Appendix A.3 for the statistics of the datasets.

4.2 Evaluation metrics

We report accuracy, precision, recall, and F1-score for ten fine-tuned claim detection models, presenting their averages and standard deviations. To account for varying ratios of verifiable and unverifiable claims across languages, we calculated weighted scores at the language level, and the overall performance is obtained by averaging the language-level performance of the models.

4.3 Baseline models

Due to the lack of cross-lingual claim detection models, we use strong cross-lingual language models as baselines: mBERT (Alam et al., 2021b; Uyangodage et al., 2021; Panda and Levitan, 2021; Zengin et al., 2021; Hasanain and Elsayed, 2022), XLM-RoBERTa (XLM-R) (Alam et al., 2021b,a), and mT5 (Du et al., 2022), all fine-tuned on the CheckThat! Train partition. These widely used models have proven competitive in multilingual claim detection. We also include open-source large language models from Li et al. (2024)—Llama2-7B, Mistral-7B, and Phi3—fine-tuned on the same data for text generation. Our comparison was restricted to open-source LLMs under 10B parameters due to the computational and resource constraints outlined in Appendix D. Refer to Appendix D for resources used for the training, hyperparameters, the prompt used for LLMs, and the LLM fine-tuning process.

5 Results

5.1 Dataset-level Performance

Table 2 summarizes the performance of the claim detection models reported as the mean and standard deviation over 10 runs. Llama2 performs strongly on the CheckThat! 2022 Test dataset, while *EXP-Claim* achieves the highest accuracy, recall, and F1-score on both the synthetic and Kazemi datasets. The performance of large language models declines on these two datasets, highlighting challenges in handling linguistically diverse data.

Notably, the EX-Claim models and their ablated variants outperform both transformer-based baselines and large language models in overall evaluation, underscoring the value of learning task-specific representations that focus on claim-relevant words. Moreover, their consistent performance across diverse datasets—spanning multiple platforms and topics—suggests that our hypothesis holds generally, regardless of the claim's nature.

Incorporating NER information (*EXN-Claim* vs. *X-Claim*) consistently improves performance across all four metrics. While the overall gain is modest (approximately 1%), our language-level analysis (Section 5.2) shows that NER integration either maintains or enhances model performance across most of the languages, with certain languages benefiting significantly from entity information and linguistic knowledge transfer from the NER and EL tools. Interestingly, incorporating en-

Model	Accuracy	Precision	Recall	F1-Score
		CheckThat! 2022		
mBERT	0.722 ± 0.006	0.723 ± 0.006	0.722 ± 0.006	0.715 ± 0.007
XLM-R	0.669 ± 0.007	0.695 ± 0.006	0.669 ± 0.007	0.658 ± 0.009
mT5	0.681 ± 0.004	0.685 ± 0.004	0.681 ± 0.004	0.679 ± 0.004
Llama2	**0.776** ± 0.001	0.782 ± 0.001	**0.776** ± 0.001	**0.771** ± 0.001
Mistral	0.769 ± 0.003	**0.784** ± 0.003	0.769 ± 0.003	0.761 ± 0.003
Phi3	0.744 ± 0.001	0.75 ± 0.001	0.744 ± 0.001	0.738 ± 0.001
X-Claim	0.747 ± 0.005	0.751 ± 0.005	0.747 ± 0.005	0.745 ± 0.005
EXN-Claim	0.754 ± 0.005	0.758 ± 0.006	0.754 ± 0.005	0.752 ± 0.006
EXP-Claim	0.755 ± 0.007	0.759 ± 0.007	0.755 ± 0.007	0.754 ± 0.007
		Synthetic Data		
mBERT	0.687 ± 0.009	0.704 ± 0.008	0.687 ± 0.009	0.679 ± 0.01
XLM-R	0.662 ± 0.008	0.68 ± 0.007	0.662 ± 0.008	0.656 ± 0.009
mT5	0.659 ± 0.007	0.662 ± 0.007	0.659 ± 0.007	0.658 ± 0.007
Llama2	0.724 ± 0.002	0.749 ± 0.002	0.724 ± 0.001	0.707 ± 0.002
Mistral	0.718 ± 0.002	**0.759** ± 0.003	0.718 ± 0.002	0.701 ± 0.003
Phi3	0.691 ± 0.002	0.735 ± 0.005	0.691 ± 0.002	0.664 ± 0.002
X-Claim	0.719 ± 0.007	0.724 ± 0.006	0.719 ± 0.007	0.718 ± 0.007
EXN-Claim	0.73 ± 0.008	0.735 ± 0.007	0.73 ± 0.008	0.729 ± 0.008
EXP-Claim	**0.731** ± 0.009	0.737 ± 0.008	**0.731** ± 0.009	**0.73** ± 0.009
		Kazemi 2021 Data		
mBERT	0.716 ± 0.007	0.725 ± 0.006	0.716 ± 0.007	0.702 ± 0.009
XLM-R	0.687 ± 0.011	0.752 ± 0.004	0.687 ± 0.011	0.686 ± 0.012
mT5	0.681 ± 0.004	0.725 ± 0.004	0.681 ± 0.004	0.678 ± 0.005
Llama2	0.633 ± 0.003	0.5 ± 0.001	0.633 ± 0.001	0.613 ± 0.004
Mistral	0.603 ± 0.004	0.747 ± 0.005	0.603 ± 0.004	0.55 ± 0.005
Phi3	0.546 ± 0.001	0.744 ± 0.001	0.546 ± 0.001	0.455 ± 0.001
X-Claim	0.788 ± 0.007	**0.817** ± 0.005	0.788 ± 0.007	0.789 ± 0.007
EXN-Claim	0.793 ± 0.008	0.813 ± 0.006	0.793 ± 0.008	0.796 ± 0.008
EXP-Claim	**0.794** ± 0.008	0.813 ± 0.006	**0.794** ± 0.008	**0.797** ± 0.007
		All		
mBERT	0.698 ± 0.008	0.711 ± 0.007	0.698 ± 0.008	0.689 ± 0.009
XLMR	0.668 ± 0.008	0.695 ± 0.006	0.668 ± 0.008	0.661 ± 0.01
mT5	0.667 ± 0.006	0.677 ± 0.006	0.667 ± 0.006	0.665 ± 0.006
Llama2	0.717 ± 0.002	0.755 ± 0.001	0.717 ± 0.002	0.702 ± 0.002
Mistral	0.707 ± 0.003	0.761 ± 0.004	0.707 ± 0.003	0.685 ± 0.003
Phi3	0.674 ± 0.002	0.74 ± 0.004	0.674 ± 0.002	0.64 ± 0.002
X-Claim	0.736 ± 0.006	0.746 ± 0.006	0.736 ± 0.006	0.735 ± 0.007
EXN-Claim	0.746 ± 0.007	0.753 ± 0.007	0.746 ± 0.007	0.745 ± 0.008
EXP-Claim	**0.747** ± 0.008	**0.754** ± 0.007	**0.747** ± 0.008	**0.746** ± 0.009

Table 2: Dataset-level Performance of the Claim Detection Models

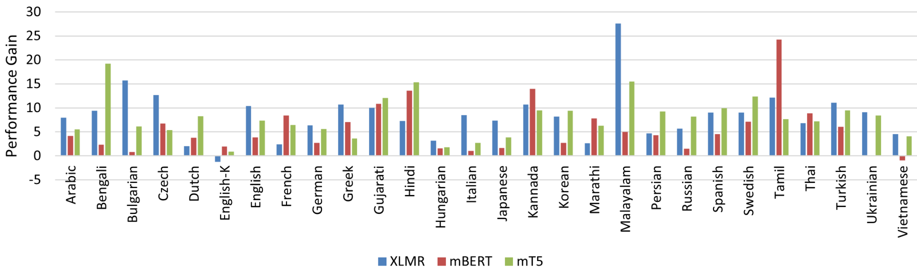
tity popularity (*EXN-Claim* vs. *EXP-Claim*) did not lead to further performance improvement—likely because most entities identified by the NER model were already recognized as popular, limiting the potential for additional gains.

5.2 Language-level Gain

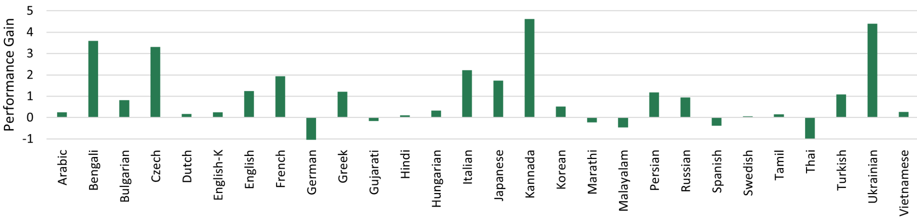
We analyze language-level performance gains from incorporating entity information into claim detection. Since there is no performance gain in *EXP-Claim* compared to *EXN-Claim*, we analyze the performance of *EXN-Claim* with three groups of models as shown in Figure 2. Refer to Appendix B.1 for the language-level F1-score of the models.

Figure 2a illustrates the performance gain of the *EXN-Claim* model over transformer-based baselines. Most languages exhibit consistent positive gains, except English (Kazemi dataset) and Vietnamese, which show slight drops (0.95–1.2%). The most substantial improvements are seen in Asian languages like Malayalam, Tamil, Hindi, Bengali, Kannada, and Gujarati. These results underscore the effectiveness of incorporating entity information and dynamically attending to claim-relevant words, leading to significant performance gains across a diverse set of languages.

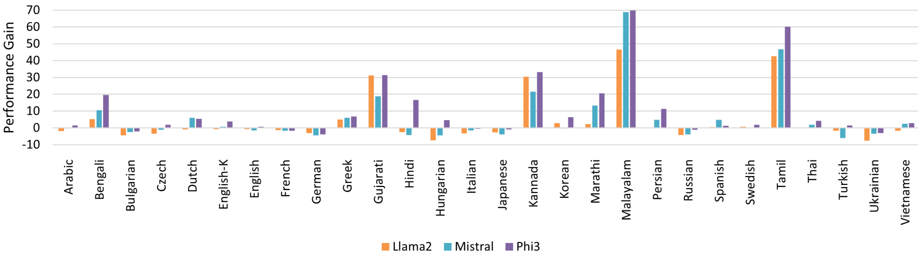
Figure 2b shows the performance gain of the *EXN-Claim* model over its ablation variant, *X-Claim*, which excludes entity information. Interestingly, incorporating entity information either retains or improves model performance across most languages. The only exceptions are German and Thai, which exhibit a slight performance drop of approximately 0.97% to 1.1%. Notably, certain languages—such as Kannada, Ukrainian, Bengali, Czech, Italian, and French—experience the high-



(a) Transformer-based Baselines



(b) *X-Claim*



(c) Large Language Models (LLMs)

Figure 2: Performance Gain in *EXN-Claim* compared to Three groups of Models

est gains, ranging from 1.95% to 4.6%. These improvements can likely be attributed to the incorporation of entity information and the indirect linguistic knowledge transfer from the NER tool.

Similarly, we compare the performance gain of the *EXN-Claim* model with large language models (LLMs) in Figure 2c. While LLMs outperform *EXN-Claim* in several high-resource languages, they perform notably poorly in certain low-resource languages—particularly Asian regional languages such as Malayalam, Kannada, Tamil, Gujarati, Bengali, and Hindi. The largest gains are in Malayalam (up to 70%) and Tamil (up to 60%), which include non-synthetic samples from the Kazemi dataset. Despite being smaller and more computationally efficient, *EXN-Claim* outperforms all three LLMs in 7 languages and surpasses at least one LLM in 16 out of the 27 languages evaluated.

Furthermore, we observe inconsistent gains across LLMs: Llama excels in Dutch, Phi3 in Hungarian, while Mistral underperforms in both. Similarly, Llama and Phi3 perform well in Korean, but Mistral struggles. These observations suggest that LLMs tend to excel only in specific languages, largely influenced by the composition of their training data. In contrast, these findings highlight *EXN-Claim* as a highly efficient model that consistently delivers superior performance across many low-resource languages, while remaining competitive with LLMs in high-resource scenarios, offering the best balance between overall performance gains and computational efficiency. Refer to Appendix D.3 for comparison of claim detection models based on the number of parameters, multilingual support, and training time.

6 Analysis

6.1 Cross-lingual Knowledge Transfer

The synthetic data comprises claims from the CheckThat! Test partition, written in 18 different languages. We use this dataset to assess the cross-lingual knowledge-transferring capabilities of the claim detection models. We compute the fraction of data instances in the Synthetic data which gets the same class prediction as its original data instance

Model	Correct Prediction	Wrong Prediction
mBERT	82.75%	64.83%
XLMR	81.72%	66.27%
mT5	79.12%	**60**%
Llama2	85.4%	70.1%
Mistral	86.87%	71.3%
Phi3	82.46%	65.5%
X-Claim	87.21%	71.73%
EXN-Claim	**87.82**%	68.48%
EXP-Claim	87.09%	66.51%

Table 3: Cross-lingual Knowledge Transfer Rate

from the CheckThat! Test partition by a model. We categorized these results into correct and incorrect predictions, where true positives and true negatives are considered correct and false positives and false negatives are treated as incorrect. This categorization allows us to evaluate the model's ability to transfer knowledge across correct and wrong predictions.

Table 3 lists the knowledge transfer rate of claim detection models. Overall, the transformer-based baseline models exhibit the lowest transfer rate in correct and wrong predictions. In contrast, *EX-Claim* and its ablated variant achieve the highest transfer rate—reaching nearly 87%—for correct predictions. This highlights the effectiveness of learning context-aware claim representations with attention to claim-relevant words in promoting cross-lingual generalization. Furthermore, the inclusion of entity information through NER and EL (*EXN-Claim* and *EXP-Claim*) reduces the transfer of incorrect predictions by 3% to 5%. This indicates that entity-level knowledge and linguistic signals from NER and EL help prevent consistent misclassification of the same instances across different languages, thus improving model robustness.

6.2 Attending to Important Words

The attention layer enables the claim detection models to attend to important words in the input sequence before generating the sentence representation. We quantify this behavior of the models as the entropy of attention weights computed by the attention layer. We compute the average entropy of the test data by computing the mean value of entropy of all the input sequences. For a fair evaluation, we sampled 250 data instances for each language in the CheckThat! 2022 and Kazemi 2021 datasets.

Table 4 lists the attention layer's entropy of the claim detection models. We can see that the models utilizing entity information always yield lower entropy, especially when utilizing popularity, resulting in the lowest entropy among all. This shows that these models are concentrated on a few keywords in the input sequence to determine the verifiability of a claim.[2]

Model	Entropy
X-Claim	4.09
EXN-Claim	3.519
EXP-Claim	**3.322**

Table 4: Entropy of Attention Layer

	False Positive
S4	*Months after receiving the second dose of the Pfizer COVID vaccine, I continue to suffer side effects including perfect health and having to deal with morons.*
S5	*How true is Congress party's claim regarding shortfall in Government earnings*
	False Negative
S6	*These are not photos of a man with an artificial heart*
S7	*Retirement Home NOT Raided By The FBI For Running Elderly Fight Club, NO 7 Arrests*
S8	*Gruesome Video Of A Man Hacked To Death In Bihar Falsely Shared As TMC 'Goons' Attacking A BJP Worker*
S9	*Video Showing Suicide Bid Falsely Linked To CAA*

Table 5: Misclassified Sample Sentences

6.3 Error Analysis

We analyze the false positive and false negative predictions made by *EX-Claim* models. Table 5 lists some of the misclassified examples. We noticed that the appearance of disease names ($S4$), lengthy claims, nonrecognition of personal experience from pronouns' presence ($S4$), and ambiguous fact expressions could cause false positive scenarios ($S5$). We observed that misclassification of popular events expressing personal experience tends to be a common cause of false negatives. Refer to Sentences *S6-S9* in Table 5 for sample false negative instances where the reach of the event makes it verifiable, yet the personal experience framing leads to misclassification.

Figure 3 presents the correlation scores between the true positive, true negative, false positive, and false negative values and selected features. It can be observed that the *length* of the claim is positively correlated with true positive and false positive rates, showing the models tend to assume lengthy claims as verifiable. Similarly, the number of disease names appearing in a claim is positively correlated with a false positive rate. This could be the COVID-

[2] Refer to Appendix C.2 for attention weights generated for sample sentences.

Text Length			
	X-Claim	EXN-Claim	EXP-Claim
True Positive	0.206	0.195	0.183
True Negative	-0.214	-0.2	-0.189
False Positive	0.098	0.085	0.072
False Negative	-0.1	-0.083	-0.063
No of. Disease Names			
	X-Claim	EXN-Claim	EXP-Claim
True Positive	-0.015	-0.014	-0.014
True Negative	-0.003	-0.005	-0.004
False Positive	0.032	0.036	0.037
False Negative	-0.009	-0.011	-0.011
No of. Media Names			
	X-Claim	EXN-Claim	EXP-Claim
True Positive	-0.089	-0.101	-0.1
True Negative	0.085	0.093	0.095
False Positive	-0.011	-0.023	-0.028
False Negative	0.021	0.037	0.034

Figure 3: Correlation of Sample Features vs Confusion Matrix

Model	F1-Score
X-Claim	**0.735**
X-Claim - NPL	0.688
X-Claim - NAL	0.714

Table 6: Ablation Results of *X-Claim* Model

19 discussion in the training data making the model to classify the claims with disease names as verifiable. Further, the number of media names is positively correlated with both true negative and false negative rates. This suggests that references to media, such as images and videos, can either enhance or hinder verifiability, depending on the context and reach.

6.4 Ablation Study

We conduct an ablation study to identify the impact of the key components of *EX-Claim* with the following variations.

- No projection layer (NPL) - The projection layer is omitted, and attention is applied directly to the embedding representations.
- No attention layer (NAL) - The attention layer is removed, hence the projected vector representations are aggregated into a sentence embedding by computing their mean.
- No embedding layer (NEL) - The embedding layer is removed, and the entity information is represented as one-hot vectors.

The *NLP* and *NAL* variations are applied to *X-Claim*, and Table 6 presents the overall F1-score of these models. The results demonstrate that the projection layer is crucial for learning task-specific input representations, as its removal leads to a 4.7% drop in overall performance. Likewise, the attention layer enhances the model's performance, contributing to a 2.1% improvement. Table 7 lists the performance of *EX-Claim* models with *NEL* variation. It can be seen, that the performance declines by approximately 1.1% - 1.3% when entity information is represented as one-hot vectors, suggesting that the models struggle to effectively utilize entity information when it is provided as a one-hot vector representation.

Model	F1-Score
EXN-Claim	**0.745**
EXN-Claim - NEL	0.733
EXP-Claim	**0.746**
EXP-Claim- - NEL	0.733

Table 7: Ablation Results of *EX-Claim Models*

7 Conclusion

We present *EX-Claim*, the first cross-lingual claim detection model that integrates entity information for improved performance in detecting claims across multiple languages. Extensive experiments on three datasets comprising 27 languages reveal that *EX-Claim* stands out as the effective model with consistent performance gain across multiple languages compared to the baselines. Language-level performance analysis reveals that certain languages—particularly Asian regional languages—benefit significantly from the incorporation of entity information. This improvement is likely driven by the emphasis on key terms essential for claim detection and by the indirect transfer of linguistic knowledge from named entity recognition and entity linking tools. Further analysis of cross-lingual transfer rates and attention layer entropy supports this finding. Notably, *EX-Claim* achieves the highest cross-lingual knowledge transfer rate across correct predictions while minimizing the wrong predictions across claims written in the same language, demonstrating its robustness in cross-lingual generalization and transfer efficiency. Future work will focus on generalizing entity-type embeddings beyond training contexts, addressing the challenges contributing to false-positive and false-negative outcomes identified in our error analysis, and systematically evaluating the approach against emerging large language models.

Limitations

The limitations of this work are as follows:

- Training data limited to a single topic - Our training data consists solely of COVID-19 discussions. Although our model generalizes well across different datasets, including those focused on political topics, this limitation may still restrict the model's broader generalization capabilities.
- Dependence on NER and entity linking tools: The proposed solution relies on existing cross-lingual NER and entity linking tools to extract entity-type information. The performance of these tools directly impacts the overall effectiveness of our model, particularly in languages where NER or entity-linking tools may perform poorly or exceptionally well.
- Training data limited to a single platform: Our training data is exclusively composed of tweets, which may cause the model to learn characteristics specific to this platform. While we demonstrate that our model can handle data from other platforms, such as WhatsApp tiplines in the Kazemi dataset, this limitation could still hinder the model's ability to generalize across different types of social media content.
- Exclusion of popular events not identified by a name: Since the system identifies and tracks the popularity of named entities, popular events not linked to any names are excluded. As previously discussed, this is a significant cause of false negative predictions.
- Bias from the use of Wikipedia as a knowledge base: We rely on Wikipedia to identify popular entities present in verifiable claims. However, it may overrepresent entities from well-documented or dominant groups, potentially introducing bias against entities associated with marginalized or underrepresented communities.

Acknowledgments

This project is funded by the European Union and UK Research and Innovation under Grant No. 101073351 as part of Marie Skłodowska-Curie Actions (MSCA Hybrid Intelligence to monitor, promote, and analyze transformations in good democracy practices). We acknowledge Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT, for enabling our experiments (King et al., 2017).

Ethical Considerations

This work relies on publicly available social media data and should be used in compliance with platform policies and privacy norms. As with other multilingual NLP systems, biases in training data and pre-trained models may lead to uneven performance across languages, particularly in low-resource settings. The proposed approach is intended to assist, not replace, human fact-checking, and its outputs should be interpreted with appropriate caution to avoid misuse or unintended societal harm.

References

S Agrestia, AS Hashemianb, and MJ Carmanc. 2022. Polimi-flatearthers at checkthat! 2022: Gpt-3 applied to claim detection. *Working Notes of CLEF*.

Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, et al. 2021a. Fighting the covid-19 infodemic in social media: a holistic perspective and a call to arms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 913–922.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, et al. 2021b. Fighting the covid-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649.

Ipek Baris-Schlicht, Angel Felipe Magnossao de Paula, and Paolo Rosso. 2021. Upv at checkthat! 2021: Mitigating cultural differences for identifying multilingual check-worthy claims. In *Proceedings of the Working Notes of CLEF 2021, Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st to 24th, 2021*, pages 465–475. CEUR.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual Autoregressive Entity Linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of* *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

SM Du, Sujatha Das Gollapalli, and See-Kiong Ng. 2022. Nus-ids at checkthat! 2022: identifying check-worthiness of tweets using checkthat5. *Working Notes of CLEF*.

Ahmet Bahadir Eyuboglu, Bahadir Altun, Mustafa Bora Arslan, Ekrem Sonmezer, and Mucahid Kutlu. 2023. Fight against misinformation on social media: Detecting attention-worthy and harmful tweets and verifiable and check-worthy claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 161–173. Springer.

Maram Hasanain and Tamer Elsayed. 2022. Cross-lingual transfer learning for check-worthy claim identification over twitter. *arXiv preprint arXiv:2211.05087*.

Wassim Henia, Oumayma Rjab, Hatem Haddad, and Chayma Fourati. 2021. icompass at nlp4if-2021–fighting the covid-19 infodemic. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 115–118.

Ahmad Hussein, Nada Ghneim, and Ammar Joukhadar. 2021. Damascusteam at nlp4if2021: Fighting the arabic covid-19 infodemic on twitter using arabert. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 93–98.

Zehra Melce Hüsünbeyi, Oliver Deck, and Tatjana Scheffler. 2022. Rub-dfl at checkthat! 2022: Transformer models and linguistic features for identifying relevant. *Working Notes of CLEF*.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*.

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott Hale. 2021. Claim matching beyond english to scale global fact-checking. 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)*, pages 4504–4517.

Thomas King, Simon Butcher, and Lukasz Zalewski. 2017. *Apocrita - High Performance Computing Cluster for Queen Mary University of London*.

Włodzimierz Lewoniewski, Piotr Stolarski, Milena Stróżyna, Elzbieta Lewańska, Aleksandra Wojewoda, Ewelina Księżniak, and Marcin Sawiński. 2024. Openfact at checkthat! 2024: Combining multiple attack methods for effective adversarial text generation.

Yufeng Li, Rrubaa Panchendrarajan, and Arkaitz Zubiaga. 2024. Factfinders at checkthat! 2024: Refining check-worthy statement detection with llms through data pruning.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, et al. 2022. Overview of the clef-2022 checkthat! lab task 1 on identifying relevant claims in tweets. In *2022 Conference and Labs of the Evaluation Forum, CLEF 2022*, pages 368–392. CEUR Workshop Proceedings (CEUR-WS.org).

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:100066.

Subhadarshi Panda and Sarah Ita Levitan. 2021. Detecting multilingual covid-19 misinformation on social media via contextualized embeddings. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–129.

Acharya Ashish Prabhakar, Salar Mohtaj, and Sebastian Möller. 2020. Claim extraction from text using transfer learning. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 297–302.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

A Savchev. 2022. Ai rational at checkthat! 2022: using transformer models for tweet classification. *Working Notes of CLEF*.

Marcin Sawiński, Krzysztof Węcel, Ewelina Księżniak, Milena Stróżyna, Włodzimierz Lewoniewski, Piotr Stolarski, and Witold Abramowicz. 2023. Openfact at checkthat! 2023: Head-to-head gpt vs. bert-a comparative study of transformers language models for the detection of check-worthy.

Kristina P Sinaga and Miin-Shen Yang. 2020. Unsupervised k-means clustering algorithm. *IEEE access*, 8:80716–80727.

PK Manan Suri and S Dudeja. 2022. Asatya at checkthat! 2022: multimodal bert for identifying claims in tweets. *Working Notes of CLEF*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Simone Tedeschi and Roberto Navigli. 2022. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.

Lasitha Uyangodage, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2021. Can multilingual transformers fight the covid-19 infodemic? In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1432–1437.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.

Muhammed Said Zengin, Yavuz Selim Kartal, and Mücahid Kutlu. 2021. Tobb etu at checkthat! 2021: Data engineering for detecting check-worthy claims. In *CLEF (Working Notes)*, pages 670–680.

A Statistics

A.1 Statistics of Named Entities Recognized

MultiNERD identifies the following 15 types of fine-grained Named Entities.

- Person (PER)
- Organization (ORG)
- Location (LOC)
- Animal (ANIM)
- Biological entity (BIO)
- Celestial Body (CEL)
- Disease (DIS)
- Event (EVE)
- Food (FOOD)
- Instrument (INST)
- Media (MEDIA)
- Plant (PLANT)
- Mythological entity (MYTH)
- Time (TIME)
- Vehicle (VEHI)

Figure 4 shows the average number of named entities recognized in the Train partition of CheckThat! 2022 data. Notably, *location* entities appear more frequently across all languages. This could be due to the COVID-19 discussion in the training data referring to the status of the disease from different countries and locations. Additionally, the occurrences of *disease*, *media*, *biological entities*, and *food* are relatively higher compared to other rare entity types, likely due to the COVID-19-related discussions present in the dataset.

A.2 Statistics of Entities Linked

Figure 5 shows the number of named entities linked to a Wikipedia page with the increase in threshold. It can be observed that the reduction in the number of entities linked to a Wikipedia page with the increase in log probability threshold is not very significant until the threshold is increased to -0.15 (probability score of 0.861). This indicates that most of the NER identified are linked to a Wikipedia page with a very high probability.

A.3 Statistics of the Datasets Used

A.3.1 CheckThat! Dataset

Table 8 presents the statistics of the CheckThat! 2022 dataset. Compared to other languages, the number of data instances is low in Dutch across all partitions except Test. Further, the ratio between verifiable and unverifiable claims is equal for Dutch languages, whereas other languages contain more verifiable statements than unverifiable (roughly 2:1 ratio).

A.3.2 Kazemi Dataset

Table 9 presents the statistics of the Kazemi dataset. The dataset contains WhatsApp tiplines and social media posts about topics such as politics and COVID-19.

B Results

B.1 Language-level Performance

Table 10 lists the language-level F1-score of the claim detection models across 27 languages present in the test datasets.

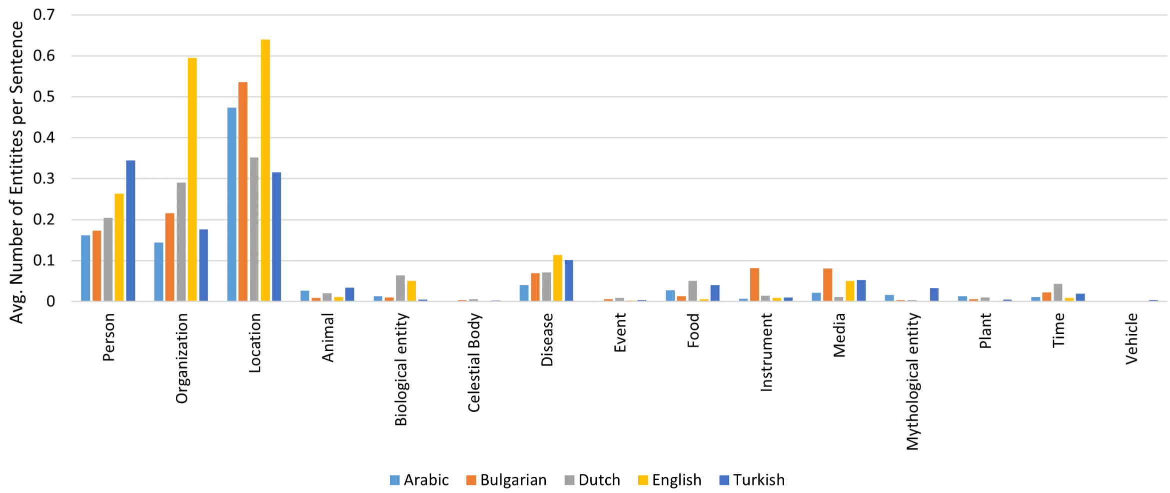


Figure 4: Average Number of Named Entities Identified in the Train Partition of CheckThat! 2022 Data

		Arabic	Bulgarian	Dutch	English	Turkish	Total
	Verifiable	2,513	1,871	929	2,122	1,589	
Train	Unverifiable	1,118	839	1,021	1,202	828	14,032
	Total	3,631	2,710	1,950	3,324	2,417	
	Verifiable	235	177	72	195	150	
Dev	Unverifiable	104	74	109	112	72	1,300
	Total	339	251	181	307	222	
	Verifiable	691	519	252	574	438	
Dev-Test	Unverifiable	305	217	282	337	222	3,837
	Total	996	736	534	911	660	
	Verifiable	682	199	608	149	303	
Test	Unverifiable	566	130	750	102	209	3,698
	Total	1,248	329	1,358	251	512	

Table 8: Statistics of the CheckThat! 2022 Verifiable Claim Detection Dataset

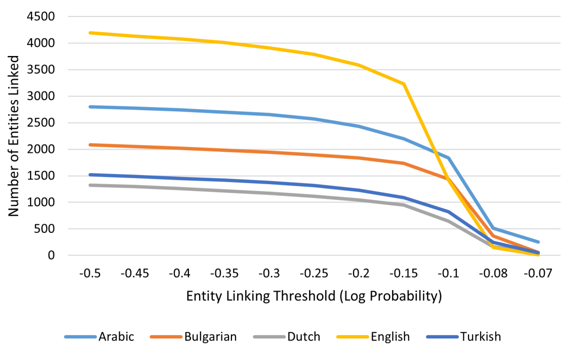


Figure 5: Number of Entities Linked with the Variation in EL Threshold in Train Partition of CheckThat! Data

B.2 Machine Translation Performance

Since the synthetic dataset was generated using machine translation, we assess the quality of the translation model—Microsoft Azure AI Translator (Junczys-Dowmunt, 2019). For this evaluation, we utilize FLORES-200 (Team et al., 2022), a multilingual benchmark dataset that supports English-to-200+ language translation. To approximate the conditions of our synthetic dataset, we sampled sentence pairs from the five source languages represented in the CheckThat! training set to 18 target languages present in the synthetic dataset. For each target language, we selected 250 translation pairs, resulting in a total of 4,500 sentence pairs. This sampling strategy is designed to mimic the structure of our synthetic dataset, where each target language includes 250 translated claims.

Metric	Score
BLEU	26.23
CHRF	54.35
CHRF++	51.63

Table 11: Machine Translation Performance of Microsoft Azure AI translator

Table 11 presents the performance of the machine translation model on the sampled translation pairs, evaluated using multiple translation quality metrics. The model achieves competitive scores on the character-based metrics CHRF and CHRF++, with the CHRF++ score notably surpassing the

	Bengali	English	Hindi	Malayalam	Tamil	Total
Verifiable	329	535	407	707	197	
Unverifiable	593	388	465	263	133	4,017
Total	922	923	872	970	330	

Table 9: Statistics of the Kazemi 2021 Verifiable Claim Detection Dataset

	Baseline						Entity Type Categories		
Language	mBERT	XLM-R	mT5	Llama2	Mistral	Phi3	X-Claim	EXN-Claim	EXP-Claim
Arabic	0.692	0.654	0.677	**0.752**	0.736	0.718	0.731	0.733	0.735
Bengali	0.725	0.655	0.556	0.697	0.642	0.552	0.713	0.748	**0.764**
Bulgarian	0.783	0.633	0.729	**0.835**	0.815	0.812	0.782	0.79	0.796
Czech	0.707	0.647	0.72	**0.808**	0.786	0.755	0.741	0.774	0.769
Dutch	0.68	0.698	0.634	0.72	0.657	0.663	0.716	0.717	**0.718**
English (Kaz)	0.738	**0.769**	0.748	0.764	0.75	0.719	0.755	0.757	0.755
English (CT)	0.724	0.658	0.688	0.769	**0.777**	0.755	0.75	**0.762**	0.761
French	0.638	0.698	0.658	0.736	**0.739**	0.738	0.703	0.722	0.725
German	0.751	0.714	0.722	0.808	**0.822**	0.816	0.789	0.778	0.775
Greek	0.677	0.64	0.71	0.697	0.687	0.679	0.735	0.747	**0.759**
Gujarati	0.661	0.669	0.649	0.457	0.582	0.455	**0.77**	**0.77**	0.751
Hindi	0.631	0.695	0.613	0.792	**0.809**	0.6	0.766	0.767	0.769
Hungarian	0.667	0.651	0.665	**0.757**	0.727	0.637	0.68	0.683	0.682
Italian	0.721	0.647	0.704	**0.764**	0.745	0.736	0.709	0.731	0.723
Japanese	0.716	0.658	0.693	0.759	**0.771**	0.74	0.715	0.732	0.746
Kannada	0.535	0.567	0.579	0.371	0.459	0.342	0.628	0.674	**0.676**
Korean	0.691	0.635	0.623	0.69	0.72	0.653	0.712	**0.718**	0.709
Marathi	0.588	0.64	0.603	0.643	0.532	0.46	**0.668**	0.666	0.664
Malayalam	0.79	0.564	0.685	0.374	0.151	0.141	**0.845**	0.84	0.835
Persian	0.673	0.669	0.622	0.713	0.667	0.602	0.704	0.716	**0.718**
Russian	0.709	0.667	0.641	**0.766**	0.764	0.734	0.715	0.724	0.724
Spanish	0.738	0.693	0.683	0.779	0.735	0.77	**0.787**	0.783	0.784
Swedish	0.707	0.688	0.654	0.772	0.779	0.76	0.777	0.778	**0.787**
Tamil	0.623	0.744	0.789	0.439	0.398	0.264	0.864	**0.866**	0.861
Thai	0.628	0.648	0.644	0.716	0.698	0.673	**0.726**	0.716	0.715
Turkish	0.699	0.648	0.664	0.775	**0.82**	0.745	0.748	0.759	0.758
Ukrainian	0.688	0.597	0.604	**0.764**	0.724	0.719	0.644	0.688	0.707
Vietnamese	0.727	0.672	0.676	**0.735**	0.694	0.689	0.715	0.717	0.729

Table 10: F1-Score of the Claim Detection Models at Language-level

baseline results reported in the original FLORES-200 dataset paper (Team et al., 2022).

C Analysis

C.1 Entity-type Embedding

Unlike the cross-lingual word embedding which is directly obtained from the pre-trained model, *EX-Claim* learns a dense vector representation for each entity type. Figure 6 shows the 2D visualization of the entity type embeddings learned by the *EXN-Claim* model. The 2D vector representation of the embeddings is obtained using the t-SNE[3] dimension reduction technique. Further, the embedding representations were clustered using the KMeans clustering technique (Sinaga and Yang, 2020), and

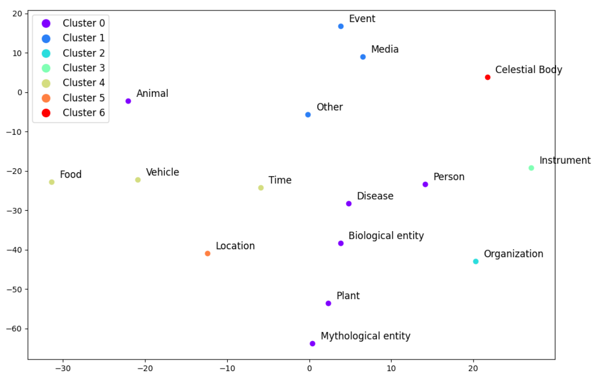


Figure 6: Two-dimensional embedding visualization of EXN-Claim Model

the number of optimal clusters was chosen using the Silhouette score (Rousseeuw, 1987).

The *EXN-Claim* model learns various groups of entity types. Notably, this model clusters *Animal,*

[3]https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

Disease, Person, Biological Entity, Disease, Plan and *Mythological Entity* into a single group, likely influenced by the COVID-19 discussions present in the training data. This observation highlights the model's tendency to learn context-specific entity-type embeddings during training.

C.2 Attention Weights for Sample Sentences

Figure 7 shows the heatmaps of the attention weights computed by different claim detection models for the following three sentences.

- *S10 - Old video of cop beating woman in **Gawalior** shared as **Delhi** police brutality*
- *S11 - **Donald Trump** fund raising email takes **CNN** anchor's comments out of context*
- *S12 - **Trump** withdraws **US** from the **Iran** nuclear deal*

It can be observed that *EX-Claim* attends to important words in the sentences, whereas *X-Claim* generates attention weights distributed across most of the words. Further, an interesting pattern of named entities attending to themselves or other named entities more compared to other words can be observed across all three sentences.

D Implementation

D.1 Resources

All the experiments were conducted using Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT (King et al., 2017). Specifically, 1 GPU (Volta V100 or Ampere A100) with 8 CPU cores, each composed of 11 GB memory was used to train and test all the models.

D.2 Models

We fine-tuned the mBERT[4], XLM-R[5], and mT5 models published in HuggingFace to obtain the transformer-based baseline models. Especially, mT5 is available in various sizes. We used mT5-large[6] as this model was the largest model among the variations which we could accommodate with the same resources utilized for training and testing other models. Similarly, the open-source large language models (LLMs) available in HuggingFace were used to obtain the LLM baselines, Llama2[7], Mistral 7B[8], and Phi3[9]. We used the same experimental setting of (Li et al., 2024). The prompt 1 used to fine-tune LLMs is a modified version of (Li et al., 2024).

Prompt 1 Verifiable Claim Detection

```
### Instruction:
Evaluate whether the input text contains
information or claims that can be verified
through fact-checking. If the statement presents
assertions, facts, or claims, respond with 'Yes'.
If the statement is purely opinion-based, trivial,
or does not contain any verifiable information or
claims, respond with 'No'.

### Input Sentence: <input sentence>

### Response: <Yes/No>
```

All the models were fine-tuned by extracting the vector representation of the special token from the final layer and applying a fully connected layer on it for classification. We used only the encoder of the mT5 model for obtaining vector representation of input words. For Named Entity Recognition (NER), we used the resources published for MultiNERD[10] for fine-tuning the fine-grained NER model. Similarly, the mGENRE[11] model published in HuggingFace was used for Entity Linking (EL).

D.3 Model Comparison

Table 12 compares claim detection models based on the number of parameters, multilingual support, and training time. Both *EXN-Claim* and EXP-Claim incur an additional average latency of 5.2 ms per sentence for named entity recognition. Furthermore, *EXP-Claim* requires an average of 0.15 seconds per entity to perform entity linking.

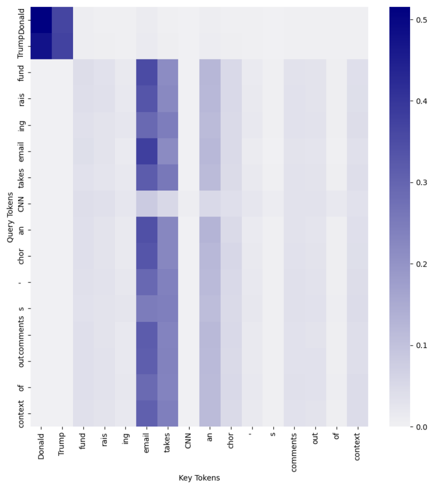
D.4 Hyperparameters

Hyperparameter	Value
Projection Size	256
Learning Rate	3e-5
Entity Embedding Size	128 (*X, EXN*), 256 (*EXP*)
Entity Linking Threshold	-0.15

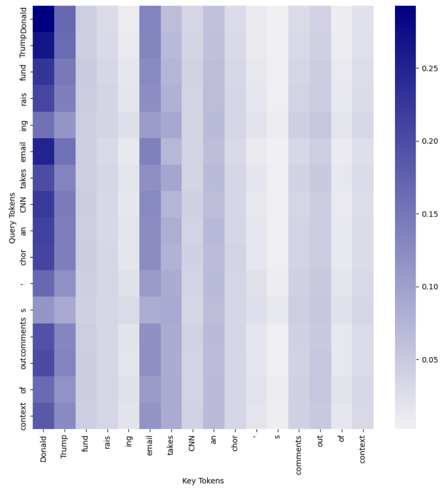
Table 13: Hyperparameters Used

Table 13 lists the hyperparameters used. Following (Devlin et al., 2019), we set the maximum length of the input sequence to 128, and use Adam as the

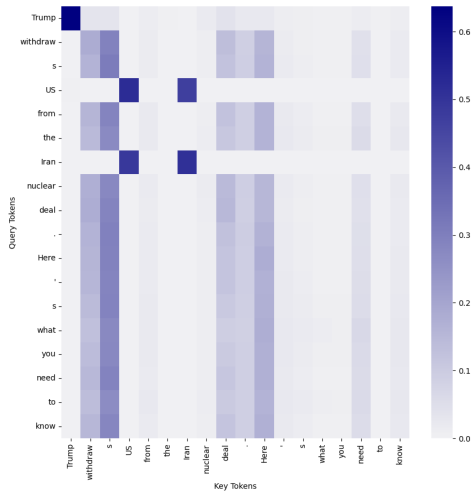
[4]https://huggingface.co/google-bert/bert-base-multilingual-cased
[5]https://huggingface.co/FacebookAI/xlm-roberta-base
[6]https://huggingface.co/google/mt5-large
[7]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[8]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
[9]https://huggingface.co/microsoft/Phi-3-mini-4k-instruct
[10]https://github.com/Babelscape/multinerd/tree/master
[11]https://huggingface.co/facebook/mgenre-wiki



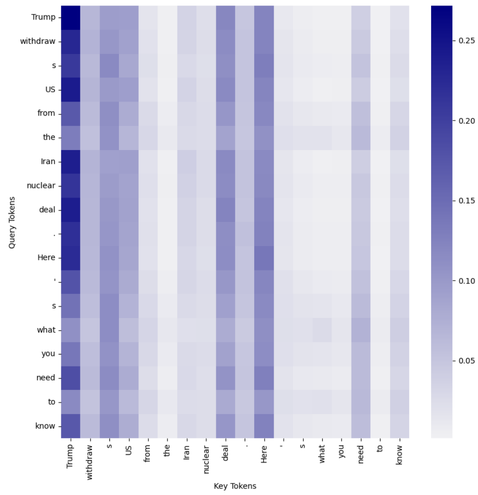
(a) S4 : *EXN-Claim*



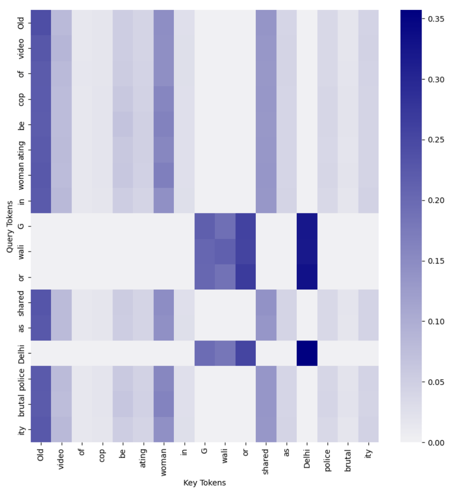
(b) S4 : *X-Claim*



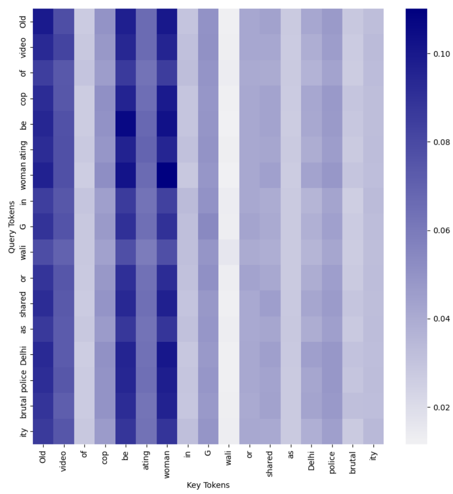
(c) S11 : *EXN-Claim*



(d) S11 : *X-Claim*



(e) S12 : *EXP-Claim*



(f) S12 : *X-Claim*

Figure 7: Heatmap of Attention Weights Generated by Claim Detection Models for sentences S10, S11, and S12

	Model	# Parameters	Multilingual?	Training time
Transformer-based Baselines	mBERT	110M	✓ Yes (104 langs)	20 sec
	XLM-R	270M	✓ Yes (100 langs)	20 sec
	mT5 Large Encoder	600M	✓ Yes (101 langs)	20 sec
LLMs	LLaMA 2	7B	Limited	3 hours & 16 mins
	Mistral	7B	Limited	3 hours & 25 mins
	Phi-3	3.8B	Limited	2 hours & 12 mins
Entity-aware Models	*X-Claim*	197K	✓ Yes (100 langs)	20 sec
	EXN-Claim	201K	✓ Yes (100 langs + NER model langs)	20 sec
	EXP-Claim	206K	✓ Yes (100 langs + NER model langs & EL model langs)	21 sec

Table 12: Comparison of Model Parameters, Multilingual Support, and Training Time

optimizer. We train the models for 30 epochs and choose the model with the lowest validation loss as the best model.