End-to-end multilingual fact-checking at scale

Vinay Setty
Factiverse AI,
vinay@factiverse.ai

Abstract

In this article, we describe how you can perform end-to-end fact-checking in over 100 languages using Factiverse AI models. We also show through an experimental benchmark that fine-tuned models tailored for fact-checking tasks outperform Large Language Models such as GPT-4, GPT-3.5-Turbo, and Mistral-7b.

1 Introduction

In the digital age, the spread of misinformation has become a significant obstacle, influencing societies, political landscapes, and public sentiments across the globe. This challenge is particularly daunting in a multilingual setting, especially when dealing with low-resource languages. Misinformation, frequently stemming from inadvertent errors by content creators, has underscored the urgent need for the creation of robust tools capable of accurately detecting and rectifying factual inaccuracies. The situation is exacerbated by the fact that online resources and tools for identifying and correcting misinformation are predominantly designed for English. This leaves speakers of low-resource languages at a significant disadvantage, highlighting a glaring gap in the global effort to combat misinformation[5]. The disparity in tool availability and effectiveness across languages poses a complex challenge, making it imperative to develop and enhance tools that cater to a broader linguistic spectrum to ensure equitable access to accurate information [3, 6].

Most newsrooms rely on content management systems for news production, offering basic formatting and composition tools. After journalists write an article, editors typically conduct manual fact-checking and proofreading, using web searches and searching internal archives. Current automation extends only to grammar checkers like Grammarly[1] and advanced tools like Writer.com[2], which automate writing styles. An early prototype as a browser plugin was developed at Factiverse which verified news articles that were already published[1]. Factiverse now offers a solution Factiverse AI Editor, an innovative text editor capable of identifying factual inaccuracies and suggesting corrections in over 100

[1] `https://www.grammarly.com`

[2] `https://writer.com`

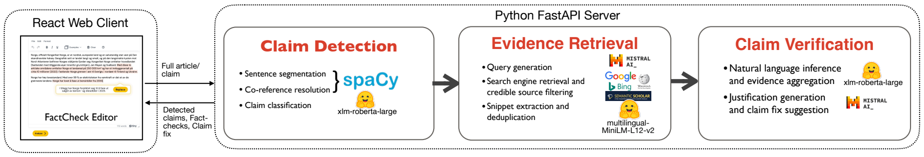

React Web Client
FactCheck Editor
Full article/ claim
Detected claims, Fact-checks, Claim fix
Python FastAPI Server
Claim Detection
• Sentence segmentation
• Co-reference resolution
• Claim classification
spaCy
xlm-roberta-large
Evidence Retrieval
• Query generation
• Search engine retrieval and credible source filtering
• Snippet extraction and deduplication
MISTRAL AI_
Google
Bing
SEMANTIC SCHOLAR
multilingual-MiniLM-L12-v2
Claim Verification
• Natural language inference and evidence aggregation
• Justification generation and claim fix suggestion
xlm-roberta-large
MISTRAL AI_

Figure 1: System Architecture of Factiverse AI Editor

languages. Factiverse AI Editor has the potential to improve content creation in sectors like news and media by helping editors detect factual errors early. However, end-to-end multilingual fact-checking presents unresolved challenges for both academia and the industry [5].

To make this problem tangible, our approach is threefold: First, we address the problem of detecting check-worthy claims within a text, a task that involves understanding the context, relevance, and potential impact of each statement. Second, we delve into the complexities of generating and executing search engine queries, which are pivotal in gathering relevant information from the web. Finally, this information is then utilized by a Natural Language Inference (NLI) model, for veracity prediction. Furthermore, we use LLMs, to generate justification summaries and also suggest precise textual amendments for error rectification.

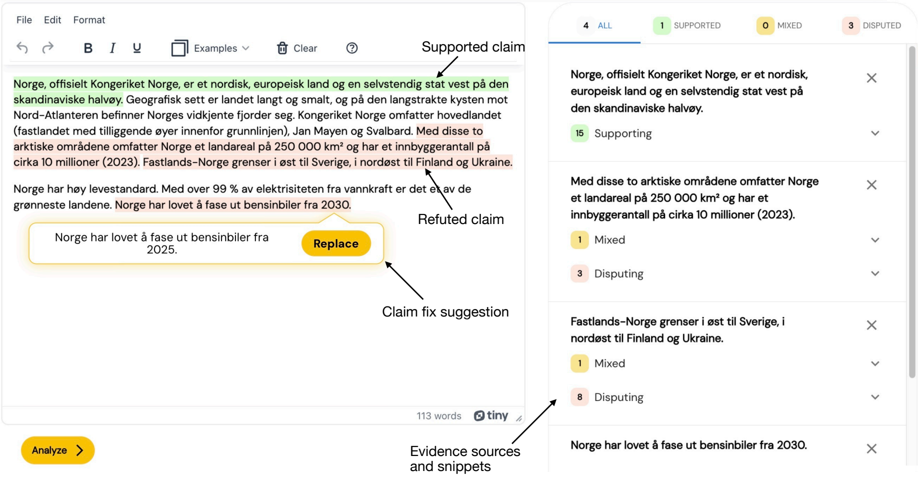
We also present preliminary evaluation results which show that a smaller multilingual model (XLM-RoBERTa-Large [2]), fine-tuned using datasets in a limited set of languages, can outperform large language models (LLMs) such as GPT-3.5-Turbo and Mistral-7b for both claim detection and veracity prediction tasks. On the other hand, LLMs excel at general tasks such as evidence summarization and suggesting corrections to false claims.

2 Factiverse AI Editor Overview

Factiverse AI Editor, is an advanced text editor designed to assist humans in productive fact-checking and facilitate correcting factual inaccuracies. Given the widespread issue of misinformation, often a result of unintentional mistakes by content creators, our tool aims to address this challenge. It supports over 100 languages and utilizes cutting-edge AI models from Factiverse to assist humans in the labor-intensive process of fact verification.

Factiverse AI Editor allows users to identify check-worthy sentences with the click of a single button (also referred to as 'claims' in the rest of the paper), in the written article by the user and verify those claims using evidence gathered from open web and previous fact-checks. Figure 1, the architecture of Factiverse AI Editor, with a web-based front-end implemented using the React framework and a backend server. The frontend includes a text editor implemented using the TinyMCE text editor[3]. The backend, exposes REST APIs to interact with

[3] https://www.tiny.cloud



Norge, offisielt Kongeriket Norge, er et nordisk, europeisk land og en selvstendig stat vest på den skandinaviske halvøy. Geografisk sett er landet langt og smalt, og på den langstrakte kysten mot Nord-Atlanteren befinner Norges vidkjente fjorder seg. Kongeriket Norge omfatter hovedlandet (fastlandet med tilliggende øyer innenfor grunnlinjen), Jan Mayen og Svalbard. Med disse to arktiske områdene omfatter Norge et landareal på 250 000 km² og har et innbyggerantall på cirka 10 millioner (2023). Fastlands-Norge grenser i øst til Sverige, i nordøst til Finland og Ukraine.

Norge har høy levestandard. Med over 99 % av elektrisiteten fra vannkraft er det et av de grønneste landene. Norge har lovet å fase ut bensinbiler fra 2030.

Norge har lovet å fase ut bensinbiler fra 2025.

Norge, offisielt Kongeriket Norge, er et nordisk, europeisk land og en selvstendig stat vest på den skandinaviske halvøy.

15 Supporting

Med disse to arktiske områdene omfatter Norge et landareal på 250 000 km² og har et innbyggerantall på cirka 10 millioner (2023).

1 Mixed

3 Disputing

Fastlands-Norge grenser i øst til Sverige, i nordøst til Finland og Ukraine.

1 Mixed

8 Disputing

Norge har lovet å fase ut bensinbiler fra 2030.

Figure 2: Factiverse AI Editor

the machine learning (ML) models. The backend is hosted on a Kubernetes cluster with autoscaling in the Google Cloud Platform. The ML models used in the backend are grouped into (a) Check-worthy claim detection, (b) Evidence search, and (c) Veracity prediction. We now explain the sub-tasks within these steps of the pipeline.

An overview of the Factiverse AI Editor usage is shown in Figure 2.

Figure 2 presents an example of a demonstration involving an article written in Norwegian that contains factual inaccuracies. For instance, the assertion "Norge har lovet å fase ut bensinbiler fra 2030" (which translates to "Norway has promised to phase out petrol cars by 2030") is flagged as incorrect, with a suggestion to replace "2030" with "2025."[4] The editor also marks claims in red and green to indicate disputed and supported claims, respectively, based on the evidence. Additionally, the right-hand pane displays evidence snippets along with a summary of the generated justification. The Factiverse AI Editor can be accessed live at `https://editor.factiverse.ai` and the evaluation code is available at `https://github.com/vinaysetty/factcheck-editor`.

3 Check-worthy Claim Detection

The goal of this stage is to quickly identify and enrich sentences in the text that warrant verification. Although there is no strict definition, there is a consensus on what constitutes a check-worthy sentence, they (1) appeal to the public to verify their correctness and veracity, and (2) do not contain subjective sentences like opinions, beliefs, or questions [5]. The first step in this stage is to identify sentences and decontextualize them to make them fully understandable

[4] `https://electrek.co/2021/09/23/norway-bans-gas-cars-in-2025-but-trends-point-toward-100-ev-sales-as-early-as`

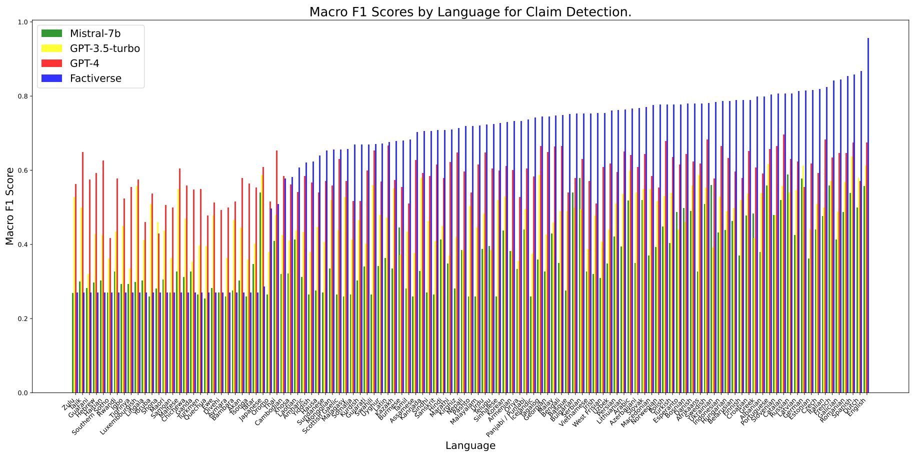


Figure 3: Evaluation of claim detection for 114 langauges using Factiverse model, GPT-3.5-Turbo, GPT-4 and Mistral-7b.

to the AI models. The sentences extracted from the text are then classified into check-worthy and not check-worthy claims using a proprietary transformer model trained using custom datasets at Factiverse. For more details, please refer to the documentation of Factiverse API for check-worthy claim detection [5].

4 Evidence Search

The goal of this stage is to search for highly relevant sources and articles that are necessary to verify the claims in the previous step. It is also important to retrieve both supporting and refuting documents for the claim. We use self-hosted LLMs like Mistral-7b to generate relevant questions and queries to search. We search Google search, Bing search, You.com, Wikipedia, Semantic Scholar (212M scholarly articles) and our own fact-checking database at Factiverse, `https://factiSearch.ai`, containing 280K fact-checks that is updated every hour. To streamline search results from various sources, we eliminate duplicates by combining URL, title, and content, using approximate matching. Additionally, we improve how relevant the information is by picking the top three paragraphs that are most closely related to the claim. We do this by using an advanced method that leverages a multilingual AI model to measure how similar the evidence snippets are to the claim, ensuring we focus on the most significant details. See here for detailed documentation of Factiverse API for evidence search: [6]

[5] `https://api.factiverse.ai/v1/redoc\#tag/Claim-Detection`

[6] `https://api.factiverse.ai/v1/redoc\#tag/Search`

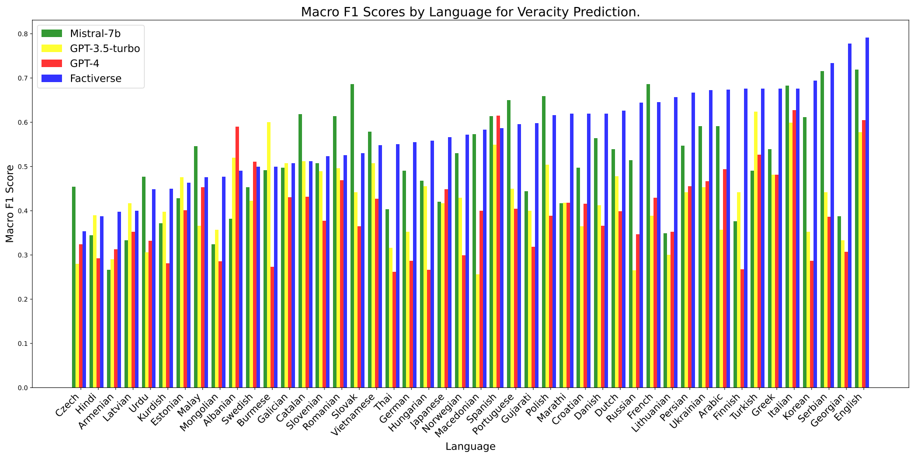


Figure 4: Evaluation of veracity prediction for 46 languages.

5 Veracity Prediction

After collecting and pre-processing the relevant evidence snippets for the identified claims, the final step in the fact-checking process is predicting the veracity based on the majority support from credible sources. We also summarize the evidence snippets found and suggest corrections to the disputed claims using an LLM. Here, we use our proprietary AI model to predict if an evidence snippet from a credible source supports or refutes the claim. We then aggregate the predictions for individual evidence snippets to come up with a verdict on if the claim is supported or disputed based on the evidence. We further summarize the results and propose a fix for the disputed claim based on the evidence using an LLM. See here for detailed documentation of Factiverse API for check-worthy veracity prediction:[7]

6 Experimental Setup

Models: We compare generative LLMs GPT-4 and GPT-3.5-Turbo models by OpenAI and Mistral-7b by Mistral AI with Factiverse models based on Transformers [4]. To adapt LLMs to perform fact-checking, we do prompt engineering to draft a prompt to predict both check-worthiness of a claim and for veracity prediction. To make the comparison fair, the same prompts are used for all LLMs. All models including Factiverse models get the identical input (claim, evidence snippets). We self-host Mistral on our servers using the ollama.ai framework.

[7] `https://api.factiverse.ai/v1/redoc\#tag/Stance-Detection`

Table 1: Dataset distribution.

Split	**Not Check-worthy**	**Check-worthy**	**True Claims**	**False Claims**	**Total**
Train	609	548	332	196	1,076
Dev	38	25	15	10	63
Test	62	38	26	12	100

Dataset: We use an internal benchmark manually by experts[8]. Since the original data is only in English, we translated the claims into 114 languages using the Google Translate API of the DeepL library[9]. An overview of the dataset is shown in Table 1.

Metrics: We use the Macro-F1 score[10], typically used to measure performance of machine learning models. Macro-F1 helps us to understand how well our models perform in scenarios when there is an imbalance between the true/false classes (applies to both claim detection and veracity prediction tasks). It does this by treating both true claims and false claims as equally important, even if there are far fewer false claims (reflects the typical scenario when fact-checking own content). This is different from accuracy, which could mislead when the data is not balanced.

If you would like to reproduce these results yourself, see the code in GitHub[11] for instructions and contact us for further details.

7 Experimental Results

7.1 Claim Detection

As shown in the Figure 3, the fine-tuned model by Factiverse impressively outperforms both OpenAI and Mistral models in most languages. Since the model was trained mainly in English, it is unsurprisingly the best-performing language. For some languages (towards the left side of the plot), we see that Factiverse is the worst-performing model. On closer inspection, these are the languages not yet supported by Factiverse. Mistral-7b seems to be the worst-performing model overall, it seems to be because Mistral struggles to follow instructions in the prompt for text classification. Table 2 shows the average Macro-F1 and Micro- F1 scores for all four models. This suggests that Factiverse models are significantly better for claim detection compared to using carefully engineered prompts with LLMs in a multilingual setting.

[8] `https://github.com/vinaysetty/factcheck-editor/tree/main/data`

[9] `https://www.deepl.com/docs-api`

[10] `https://en.wikipedia.org/wiki/F-score`

[11] `https://github.com/vinaysetty/factcheck-editor`

Table 2: Claim detection and veracity prediction results presented as mean Micro and Macro-F1 scores for all languages.

Model	**Claim Detection**		**Veracity Prediction**	
	Ma.-F1	**Mi.-F1**	**Ma.-F1**	**Mi.-F1**
GPT-4	0.624	0.591	0.460	0.426
GPT-3.5-Turbo	0.562	0.567	0.440	0.396
Mistral-7b	0.477	0.510	0.509	0.557
Factiverse	**0.743**	**0.768**	**0.575**	**0.594**

7.2 Veracity Prediction

As shown in the Figure, 4, the fine-tuned Factiverse model outperforms the other models in 37 languages. GPT-4 is the best model only for three languages: Swedish, Albanian, and Georgian. Mistral-7b is the best model in 8 languages, and it is interesting to see that Mistral performs better than GPT-3.5-Turbo/GPT-4 models despite being a much smaller LLM. Mistral-7b seems to be the best model for some European languages, such as French, Spanish, Catalan, and Portuguese. Since, for some languages, we couldn't find any evidence snippets from the search engines for any of the claims, they are omitted. Overall results for the veracity prediction are summarized in Table 2.

8 Conclusion

In this article, we showed that fine-tuned models used at Factiverse, even if they are much smaller, can outperform LLMs in a multilingual setting. However, there is some room for improvement for Factiverse models in some European languages, such as French, Spanish, and Portuguese.

References

[1] B. Botnevik, E. Sakariassen, and V. Setty. Brenda: Browser extension for fake news detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pages 2117–2120.

[2] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.

[3] Z. Guo, M. Schlichtkrull, and A. Vlachos. A survey on automated fact-checking. 10:178–206.

[4] R. Mishra and V. Setty. Sadhan: Hierarchical attention networks to learn latent aspect embeddings for fake news detection. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 197–204, New York, NY, USA, 2019. Association for Computing Machinery.

[5] R. Panchendrarajan and A. Zubiaga. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research.

[6] X. Zhou and R. Zafarani. Fake news: A survey of research, detection methods, and opportunities.