# Cross-Lingual Learning vs. Low-Resource Fine-Tuning: A Case Study with Fact-Checking in Turkish

**Recep Firat Cekinel**[1,2], **Pinar Karagoz**[1], **Cagri Coltekin**[2]
[1]Middle East Technical University
Ankara, Turkiye
{rfcekinel, karagoz}@ceng.metu.edu.tr

[2]University of Tübingen
Tübingen, Germany
ccoltekin@sfs.uni-tuebingen.de

**Abstract**

The rapid spread of misinformation through social media platforms has raised concerns regarding its impact on public opinion. While misinformation is prevalent in other languages, the majority of research in this field has concentrated on the English language. Hence, there is a scarcity of datasets for other languages, including Turkish. To address this concern, we have introduced the FCTR dataset, consisting of 3238 real-world claims. This dataset spans multiple domains and incorporates evidence collected from three Turkish fact-checking organizations. Additionally, we aim to assess the effectiveness of cross-lingual transfer learning for low-resource languages, with a particular focus on Turkish. We demonstrate in-context learning (zero-shot and few-shot) performance of large language models in this context. The experimental results indicate that the dataset has the potential to advance research in the Turkish language.

**Keywords:** misinformation, fact-checking, cross-lingual learning

## 1. Introduction

Progresses in social networking and social media have not only made information more accessible but have also enabled the rapid spread of false information on these platforms (Vosoughi et al., 2018). As a result, disseminating fake stories has emerged as a powerful instrument for manipulating public opinion, as observed during the 2016 US Presidential Election and the Brexit referendum (Pogue, 2017; Allcott and Gentzkow, 2017). Fake news can be described as media content that contains false information with the intent to mislead individuals (Shu et al., 2017; Zhou and Zafarani, 2020). The goal of fake news detection is to evaluate the correctness of statements within the message content.

The traditional method of evaluating the correctness of a claim involves seeking the expertise of specialists who assess the claim by examining the available evidence. For instance, organizations like PolitiFact[1] and Snopes[2] rely on editors to validate the correctness of statements. However, this approach is both time-consuming and expensive. To address this issue, automated methods for fact-checking have emerged, intending to assess the truthfulness of claims while reducing the need for human intervention (Oshikawa et al., 2020).

Like many other problems in NLP, the vast majority of available fact-checking resources released are primarily in English (Guo et al., 2022). However, misinformation is not specific to content generated in English. Automated fact-checking systems are also needed for other languages, despite having much lower amount of expert annotated fact-checking data. Besides supervised data availability, the distribution of languages in pretraining data of state-of-the-art models also creates a big imbalance between English and other languages. Since creating large, manually annotated fact-checking data is a very expensive endeavor, and finding the amount of unannotated data in languages other than English to (pre)train large language models are impractical (if not impossible), one promising solution is linguistic transfer: leveraging large datasets in English and cross-lingual transfer learning methods to build fact-checking systems for other, low-resource languages.

Cross-lingual learning has been studied in related problems such as hate speech detection (Stappen et al., 2020), rumor detection (Lin et al., 2023), abusive language detection (Glavaš et al., 2020) and malicious activity detection on social media (Haider et al., 2023). For fact-checking, Du et al. (2021) proposed a model that jointly encodes COVID-19-related Chinese and English texts. Additionally, Raja et al. (2023) employed joint training of English and Dravidian news articles and also applied zero-shot transfer learning by fine-tuning with English data and testing on Dravidian data.

Our primary aim in this study to test the viability of cross-lingual transfer learning approaches

[1]https://www.politifact.com/
[2]https://www.snopes.com/fact-check/



*LREC-COLING 2024*, pages 4127–4142

for fact-checking. We particularly focus on making use of data in English for fact-checking in Turkish for the cases of no or limited data availability. For this purpose, we collect a fact-checking data set for Turkish, and perform experiments with transfer learning through fine-tuning large language models and utilizing machine translation. Besides an assessment of the feasibility of transfer learning approaches, our results also provide some preliminary evidence for the type of information, knowledge or style, used in automated fact-checking models.

Our contributions can be summarized as:

- Releasing a Turkish fact-checking dataset obtained by crawling three Turkish fact-checking websites.[3]
- Assessing the efficiency of transfer learning for low-resource languages, with a specific emphasis on Turkish.
- Presenting experimental results, comparing zero- and few-shot prompt learning and fine-tuning on large language models and underscoring the need to utilize a small amount of native data.

## 2. Related Work

**Datasets.** In recent years, numerous datasets have emerged for fact-checking and they can be categorized based on how claim statements are obtained. Some studies that create claim statements by extracting and manipulating content from source documents such as Wikipedia articles can be categorized as artificial claims (Thorne et al., 2018; Jiang et al., 2020; Schuster et al., 2021; Aly et al., 2021; Kim et al., 2023). These studies involve human annotators who systematically generate meaningful claims.

On the other hand, another approach involves collecting claims by crawling fact-checking websites such as Politifact (Vlachos and Riedel, 2014; Wang, 2017) that primarily focuses on political claims and Snopes (Hanselowski et al., 2019) that covers a broader range of topics. Additionally, some studies gather fact-checked claims from the Web (Augenstein et al., 2019; Khan et al., 2022), specifically targeting domains like healthcare (Kotonya and Toni, 2020b; Sarrouti et al., 2021), science (Wadden et al., 2020), e-commerce (Zhang et al., 2020). Furthermore, Su et al. (2023) introduced a hybrid dataset that includes both human-annotated and language model-generated claims.

Fact-checking datasets in languages other than English, and multilingual datasets are limited in comparison to English. FakeCovid (Shahi and Nandini, 2020) includes 5182 multilingual news articles related to COVID-19. DANFEVER (Nørregaard and Derczynski, 2021), a Danish fact-checking dataset, comprises 6407 claims generated systematically following the FEVER (Thorne et al., 2018) approach. Similarly, CsFEVER (Ullrich et al., 2023) features 3097 claims in Czech using a similar methodology. Additionally, CHEF (Hu et al., 2022) contains 10K claims in Chinese. Furthermore, CT-FCC-18 (Barrón-Cedeno et al., 2018) contains political fact-checking claims in both English and Arabic, focusing on the 2016 US Election Campaign debates. X-Fact (Gupta and Srikumar, 2021) comprises 31189 short statements from fact-checking websites across 25 languages. Lastly, Dravidian_Fake (Raja et al., 2023) consists of 26K news articles in four Dravidian languages.

The majority of existing datasets have concentrated on textual content for fact-checking. Nevertheless, some claims can benefit from the integration of various modalities, including images, videos and audio. Resende et al. (2019) provides video, image, audio and text content from WhatsApp chats to detect the dissemination of misinformation in Portuguese. Nakamura et al. (2020); Luo et al. (2021); Abdelnabi et al. (2022); Yao et al. (2023); Suryavardan et al. (2023) utilize both visual and textual information for fact-checking. Additionally, MuMiN (Nielsen and McConville, 2022) incorporates the social context in the X platform (aka Twitter) and includes 12914 claims in 41 languages.

To the best of our knowledge, the only other fact-checking dataset that includes Turkish is X-Fact (Gupta and Srikumar, 2021) which includes claims and evidence documents in 25 languages. Besides the differences in size of the corpus, their Turkish data diverges from ours in a number of ways. Mainly, our focus in the corpus collection is richer monolingual data, rather than a large coverage of languages. The evidence documents in X-fact are through web searches, rather than crawling directly from the fact-checking site. Although there is some overlap in our sources, our data is also more varied in terms of fact-checking sites and topics of the claims. We also include short summaries provided in justifications and additional metadata. The summaries can be valuable for explainability in fact-checking (Atanasova et al., 2020a; Kotonya and Toni, 2020b; Stammbach and Ash, 2020; Brand et al., 2022; Cekinel and Karagoz, 2024). In addition, a semi-automated method is applied to eliminate duplicate claims that we crawled from different sources.

[3] https://github.com/firatcekinel/FCTR

**Methods.** Automated fact-checking has been studied from data mining (Shu et al., 2017) and natural language processing (Oshikawa et al., 2020; Guo et al., 2022; Vladika and Matthes, 2023) perspectives. The methods can be classified as content-based and context-based.

Zhou and Zafarani (2020) further classify content-based methods as knowledge-based (Pan et al., 2018; Cui et al., 2020) and style-based (Zhou et al., 2020; Pérez-Rosas et al., 2018; Jin et al., 2016; Jwa et al., 2019). Both approaches utilize news content to verify the veracity of a statement. While knowledge-based models assess statements by referencing their knowledge base, style-based methods typically prioritize assessing the lexical, syntactic and semantic attributes during verification.

Similarly, the authors categorized context-based methods as propagation-based (Hartmann et al., 2019; Zhou and Zafarani, 2019) and source-based (Sitaula et al., 2020). Both methods aim to capture social context to uncover the spread of information. While propagation-based models leverage interactions among users on social media by enhancing the interaction network with additional details like spreaders and publishers, source-based approaches rely on the credibility of sources which can also be employed to identify bot accounts on social media.

Kotonya and Toni (2020a) conducted a survey of the explainable fact-checking literature and classified the studies based on explanation generation approaches.These methods include exploiting neural network artifacts (Popat et al., 2017, 2018; Shu et al., 2019; Lu and Li, 2020; Silva et al., 2021), rule-based approaches (Szczepański et al., 2021; Gad-Elrab et al., 2019; Ahmadi et al., 2020), summary generation (Atanasova et al., 2020a; Kotonya and Toni, 2020b; Stammbach and Ash, 2020; Brand et al., 2022; Cekinel and Karagoz, 2024), adversarial text generation (Thorne et al., 2019; Atanasova et al., 2020b; Dai et al., 2022), causal inference (Cheng et al., 2021; Zhang et al., 2022; Li et al., 2023; Xu et al., 2023), neurosymbolic reasoning (Pan et al., 2023; Wang and Shu, 2023) and question-answering (Ousidhoum et al., 2022; Yang et al., 2022).

Transfer learning approaches are relatively rare for fact-checking. One approach in this field focuses on claim matching, aiming to link a claim in one language with its fact-checked counterpart in another language (Kazemi et al., 2021, 2022). Another approach focuses on out-of-domain generalization, involving the training of multilingual language models in a cross-lingual context (Gupta and Srikumar, 2021). Besides, cross-lingual evidence retrievers can be employed to retrieve evidence documents in any language corresponding to a claim made in a different language, thereby enhancing the cross-lingual fact-checking capabilities (Huang et al., 2022).

# 3. Data

Fact-checking datasets in both Turkish and English, are released by crawling Turkish fact-checking organizations and Snopes for English content. The significant similarity between the fact-checking domains of the Turkish websites and Snopes presents a valuable opportunity for transfer learning. In this study, various experiments are conducted to evaluate the necessity of collecting datasets in low-resource languages versus the effectiveness of transfer learning for these languages. Furthermore, we also conducted topic modeling to explore the latent topics within the datasets in Appendix A and examined the potential content-based discrepancies between true and fake claims in Appendix B.

## 3.1. Dataset for Fact-Checking in Turkish (FCTR)

We crawled 6787 claims from the three Turkish fact-checking websites: Teyit, Dogrulukpayi and Dogrula.[4] All are listed as fact-checking organizations on the Duke Reporters' Lab.[5] Dogrulukpayi and Teyit are also members of the International Fact-Checking Network (IFCN) which is a global community of fact-checkers. Our data collection process involved extracting *claim statements*, the corresponding *evidence* presented by the editorial teams, *summaries* providing justifications which are also written by the editors, *veracity labels*, *website URLs* and the *publication dates* of the URLs.

Claims retrieved from Teyit are summarized using the 'findings' section, which provides an overview of the evidence statements. Likewise, when it comes to claims sourced from Dogrula, the summary is derived from the final paragraph within the 'evidences' section, encapsulating the key findings. In the case of claims obtained from Dogrulukpayi, the dataset includes a dedicated paragraph following the rating section that encapsulates both the claim and the supporting evidence. This paragraph serves as the summary of these claims. Moreover, unique IDs were assigned to each claim in the dataset.

Claims were also marked as multi-modal if they contained keywords such as 'video', 'photo' and 'image' etc. This classification was made because

[4] https://teyit.org/analiz, https://www.dogrulukpayi.com, https://www.dogrula.org/dogrulamalar

[5] https://reporterslab.org/fact-checking/

Figure 1: A fact-checked claim with multi-modal components [7]

| Label | Sources | Counts |
|---|---|---|
| false | Dogrula, Teyit, Dogrulukpayi | 2780 |
| true | Dogrula, Teyit, Dogrulukpayi | 203 |
| mixed | Teyit | 109 |
| partially false | Dogrulukpayi | 72 |
| unproven | Teyit | 37 |
| half true | Dogrula | 17 |
| mostly false | Dogrula | 14 |
| mostly true | Dogrula | 6 |

Table 1: Veracity label counts in the FCTR dataset

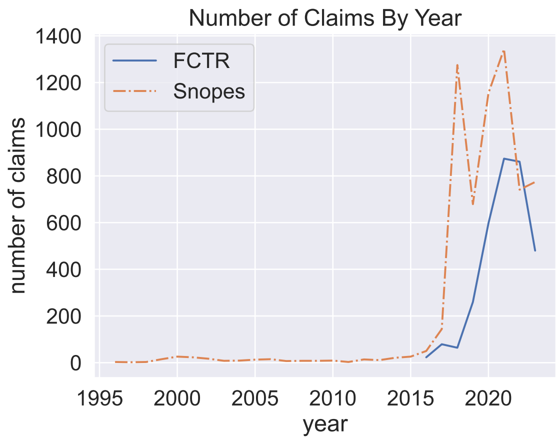


Figure 2: Number of claims by year in FCTR and Snopes datasets

we recognize that claims featuring such terms require verification not only of their textual content but also of any associated visual or video elements. For example, consider the fact-checked claim presented in Figure 1, which includes an image. In this claim, it was stated that the video shared on social media shows the moments when protesters in France set fire to the Alcazar Library in Marseille during the recent protests. The reviewer who gathered supporting information noted that 'According to inverse visual search results, the video is not from Marseille; it's from the Philippines. The building that caught fire is the Manila Central Post Office.' As a result, in order to verify such claims every aspect of evidences should be processed. Since our focus in this study is linguistic aspects of fact-checking, we do not make use of claims that require multimodal processing.

Last but not least, since the claims were collected from three distinct sources, we reviewed the claims to identify candidate duplicate claims. To accomplish this, the BERTScore metric (Zhang et al., 2019) was employed that calculates a similarity score by analyzing the contextual embeddings of individual tokens within claim statements. We set the similarity threshold to 0.85 and execute the metric three times in data source pairs. Subsequently, a manual verification process was conducted to confirm whether the outputs from BERTScore indeed corresponded to duplicate claims.

After the preprocessing step, the dataset contains 3238 claims dating from July 23, 2016 to July 11, 2023. The value counts for each label are presented in Table 1. Furthermore, 742 claims of the final dataset were sourced from Dogrulukpayi, 525 claims were retrieved from Dogrula and 1971 fact-checked claims were gathered from Teyit.

### 3.2. Snopes Dataset

Snopes is an independent organization committed to fact-checking in English. They employ human reviewers who collect information about claims and write detailed explanations as justifications. It covers a broad range of topics, including politics, health, science, popular culture, etc. We collected claims along with their metadata including the justifications written by human annotators, veracity labels, website URLs and publication dates. We collected 6402 claims ranging from November 24, 1996 to August 17, 2023 and the label distribution is shown in Table 2. Even though Snopes covers a significantly wider date range than the FCTR, the majority of claims are verified within the period from 2015 to 2023 as illustrated in Figure 2.

To the best of our knowledge, Snopes corpus was also crawled by Hanselowski et al. (2019); Augenstein et al. (2019). The reason why we recollected the Snopes claims is that the previous corpus were released in 2019 but our FCTR corpus is up-to-date. Since we aim to evaluate the effectiveness of cross-lingual transfer learning and considering the potential overlap in fact-checking

[7] https://teyit.org/analiz/videodaki-yanginin-marsilyadaki-kutuphaneden-oldugu-iddiasi

[9] 'other' encompasses the following labels: scam, outdated, misattributed, originated as satire, legend, research in progress, fake, recall, unfounded, legit

| Veracity Labels | Counts |
|---|---|
| false | 2270 |
| true | 1467 |
| mixture | 588 |
| miscaptioned | 375 |
| unproven | 284 |
| labeled satire | 283 |
| correct attribution | 247 |
| mostly false | 237 |
| mostly true | 198 |
| other | 453 |

Table 2: Veracity label counts in the Snopes dataset[9]

similar claims across both languages, we gathered the recent fact-checked claims in both English and Turkish.

## 4. Method

**Model.** In this study, we fine-tuned the LLaMA-2 (Touvron et al., 2023) model for the veracity prediction task. Llama-2 is an open-source, autoregressive transformer-based language model that was released by the Meta AI team. It has three variants, with parameter sizes of 7 billion, 13 billion, and 70 billion. Our main rationale for utilizing Llama-2 is that it has a very large and almost up-to-date knowledge base. To be more specific, the pretraining data includes information up to September 2022, while the fine-tuning data is up to June 2023.

State-of-the-art language models comprise billions of parameters, demanding large GPU memory resources during fine-tuning for downstream tasks. Additionally, the deployment of such models in real-time applications has become increasingly impractical. Therefore, we adopted parameter-efficient fine-tuning and quantization to make the Llama-2 model fit within our GPU memory constraints without sacrificing information. First, LoRA (Hu et al., 2021) introduces a small number of additional parameters and updates their weights while keeping the original parameters frozen. Similarly, QLora (Dettmers et al., 2023) employs quantization to the frozen parameters to increase memory efficiency without a significant trade-off.

**Instruction Prompting.** Instruction tuning is a method that involves additional training of language models using template instruction-output pairs. It is shown that instruction tuning significantly improves the performance of large language models across a range of tasks (Zhang et al., 2023). This is because feeding such tuples to describe the task, allows it to better grasp the domain in question. Additionally, prompting was shown to be an effective way to describe models' reasoning steps by enabling the generation of coherent reasoning chains leading to the desired output (Wei et al., 2022).

Zero-shot prompting is a method of instructing a language model to generate predictions based on a provided prompt template, without the need for specific examples. During this decision-making process, language models can utilize both the knowledge that they acquired during pretraining and the template prompt. Zero-shot prompting proves particularly useful when you have fine-tuned a language model for a related task but lack labeled data for the specific task at hand. On the other hand, providing one or more examples from the intended task as prompts is referred to as few-shot prompting. By presenting these samples within the prompt, the model gains a better understanding of the desired output and its structure. Therefore, it often leads to superior performance compared to zero-shot prompting.

## 5. Experiments and Results

This section assesses the efficacy of transfer learning in the context of low-resource languages with a specific focus on Turkish. Note that only the best results achieved during the validation experiments for each model are presented.

### 5.1. Setup

The experiments were performed on two distinct datasets: *Snopes* and *FCTR*. Given the highly imbalanced nature of the Turkish fact-checking dataset, we conducted experiments on two variants of *FCTR*, namely *FCTR500* and *FCTR1000*. In the *FCTR500* dataset, all true claims along with 297 randomly sampled false claims were included. Conversely, in the *FCTR1000* dataset, 797 false claims were randomly sampled and combined with 203 true claims. *FCTR500* represents a balanced dataset, while *FCTR1000* serves as its imbalanced counterpart. Other labels were excluded because of their relatively low instance count and the varying labeling conventions within fact-checking communities for ambiguous cases such as partially true and unproven claims. Similarly, when evaluating the language models on the Snopes dataset, we focused specifically on true and false instances. In both datasets, we randomly select 80% of the data for training, 10% for validation, and 10% for testing.

The SVM model (Cortes and Vapnik, 1995) and the multilingual BERT (mBERT) model (Devlin et al., 2019) were both trained on the same datasets with identical train-dev-test partitions as

```
### Instruction: Is the following statement "true" or "false"?
### Input:
A series of photographs show the skeletal remains of the biblical giant Goliath.
### Response:
false
```

Figure 3: Prompt template

a baseline. For the SVM model, we used sparse word and n-gram features weighted by tf-idf. The training instances are weighted with inverse class frequency to counteract the class imbalance, particularly in the case of *FCTR100* trials. Similarly, we modified the cross-entropy loss function for the mBERT model. This adaptation took into account the inverse class ratios, causing the models to assign a higher penalty to the errors on the minority class compared to the majority class.

Prompt engineering played a critical role in the experiments. Various prompt formats were evaluated and the best results were achieved using the Alpaca prompt template (Taori et al., 2023), which is provided in Figure 3. The LLaMA-2 implementations in the Huggingface's transformers library[10] were utilized language models in our transfer learning experiments. Although the LLaMA-2 language model was primarily pretrained on English data, we confirmed its proficiency in Turkish as well. Since it was pretrained on relatively recent data, we preferred LLaMA-2 in our experiments.

In the experiments, we used the SFTTrainer (from trl library) to fine-tune our models. While fine-tuning the LLMs cross entropy loss and Adam optimizer (paged_adamw_32bit) with linear scheduler were employed. Additionally, we used a half-precision floating point format (fp16) to accelerate computations. Moreover, we applied parameter-efficient fine-tuning utilizing the QLoRA (Dettmers et al., 2023) method to fit the language models to Nvidia Quadro RTX 5000 and Nvidia RTX A6000 GPUs. The configuration included setting the dimension of the low-rank matrices (r) to 16, establishing the scaling factor for the weight matrices (lora_alpha) at 64, and specifying a dropout probability of 0.1 for the LoRA layers (lora_dropout).

## 5.2. Evaluation

In its prototypical use, fact-checking is very similar to many retrieval problems. We would like to identify a few non-factual texts (e.g., fake news) among (presumably) many factual documents (legitimate news). As a result, binary precision, recall and F1 scores considering non-factual texts as positive instances is a natural choice for evaluation. However, the datasets at hand provide an interesting challenge for evaluating fact-checking models. Since both classes are obtained from fact-checking organizations, most claims they care to consider are not factual.[11] Hence, the data sets at hand show a reverse class-imbalance compared to what we expect to observe in real use of such systems. As a result, for all experiments reported in this paper, we report F1-macro and F1-binary scores with respect to the 'false' class. The hyper-parameter sweeps are performed to optimize the F1-macro score.

| Input | Model | F1-macro | F1-binary |
|---|---|---|---|
| claim 10-fold | SVM | 0.651 | 0.709 |
| claim | SVM | 0.695 | 0.763 |
| claim | mBERT | 0.705 | 0.802 |
| claim | LLaMA-7B | 0.766 | 0.838 |
| claim | LLaMA-13B | 0.814 | 0.866 |
| claim | LLaMA-70B | **0.826** | **0.890** |

Table 3: Veracity prediction on the Snopes data

## 5.3. Results

**Snopes Results.** First of all, we conducted fine-tuning of the LLaMA and baseline models using the Snopes dataset. In all trials, input consisted solely of claim statements, without the inclusion of any supporting evidence. The results are summarized in Table 3. According to the results, the LLaMA-2 model with 70 billion parameters exhibited the best performance compared to other models. Since no supporting evidence was provided, the models were expected to rely on stylistic features for their predictions. It is noteworthy that the SVM models learned purely from stylistic features. Nevertheless, a substantial performance gap exists between the SVM and the LLaMA-2 models. This margin could be attributed to the pretrained knowledge embedded in LLaMA-2 models. Moreover, the larger LLaMA-2 models outperformed LLaMA-7B, suggesting that LLaMA-13B and LLaMA-70B leverage their knowledge better than their smaller variant.

[10] https://huggingface.co/meta-llama

[11] Obtaining claims by other means may be a possible way to restore the class balance. However, such an approach also risks introducing spurious correlations with the veracity label (e.g., topic, style due to collection procedure).

| Input | Model | F1-macro | F1-binary |
|---|---|---|---|
| claim 10-fold | SVM | 0.682 | 0.610 |
| claim | SVM | **0.714** | 0.709 |
| claim | mBERT | 0.653 | 0.750 |
| claim | LLaMA-7B | 0.632 | 0.765 |
| claim | LLaMA-13B | 0.635 | 0.679 |
| claim | LLaMA-70B | 0.649 | **0.783** |
| +summary | mBERT | 0.752 | 0.861 |
| +summary | LLaMA-13B | **0.890** | **0.923** |

Table 4: Fine tuning on the FCTR500 data

| Input | Model | F1-macro | F1-binary |
|---|---|---|---|
| claim | SVM | **0.671** | 0.842 |
| claim | mBERT | 0.518 | 0.797 |
| claim | LLaMA-7B | 0.561 | **0.864** |
| claim | LLaMA-13B | 0.642 | 0.839 |
| +summary | mBERT | 0.729 | 0.902 |
| +summary | LLaMA-13B | **0.828** | **0.947** |

Table 5: Fine tuning on the FCTR1000 data

**FCTR Results.** Table 4 and Table 5 present the fine-tuning results on the *FCTR500* and *FCTR1000* datasets respectively. According to the findings, when using only the claim statement as input, the SVM model which bases its predictions solely on stylistic features achieved the highest F1-macro score on the *FCTR500* and *FCTR1000* datasets. While evaluating with claim statements only, on *FCTR1000* dataset, we fine-tuned the LLaMA models on the Snopes dataset for two epochs initially and continued fine-tuning on the *FCTR1000* dataset for one epoch to achieve the best results. Besides, the class weights of the cross entropy loss function of the multilingual BERT model were adjusted according to the class proportions inversely to get the best result.

Furthermore, when both the claim statement and the summary (which summarizes the evidence provided by crowd workers) were given as input, the LLaMA-13B model reached a superior 0.89 and 0.828 F1-macro scores on *FCTR500* and *FCTR1000* datasets respectively and 0.923 and 0.947 F1-binary scores respectively. These scores were substantially higher compared to training the model with claims alone. The reason why we incorporated summaries as input was to examine whether this additional information improves the models' capabilities. Notably, the LLaMA models have limited proficiency in Turkish and we observed poor performance when solely presented with claim statements.

**Assessing the Impact of Number of Training Instances.** In this experiment, we examined the influence of varying training data quantities on model performance. We maintained consistency by utilizing the identical test set employed in the previous experiment given in Table 4. Table 6 illustrates the consequences of manipulating the quantity of training data when employing the LLaMA-7B model. According to the results, as the number of training instances increases, the F1-macro score exhibits gradual improvement. However, when we employed 300 and 400 training instances, the model's performance remained almost constant, with both cases yielding remarkably similar results with only a single instance having a label change in the negative direction. This observation suggests that beyond a certain threshold, additional training instances may not provide substantial performance gains, highlighting the presence of a saturation point in the learning curve.

| Model | Input | F1-macro | F1-binary |
|---|---|---|---|
| LLaMA-7B | 50 claims | 0.566 | 0.644 |
| LLaMA-7B | 100 claims | 0.570 | 0.716 |
| LLaMA-7B | 200 claims | 0.576 | 0.677 |
| LLaMA-7B | 300 claims | **0.649** | **0.783** |
| LLaMA-7B | 400 claims | 0.632 | 0.765 |

Table 6: Impact of number of inputs on the FCTR500 data

## 5.4. Cross-Lingual Transfer Learning

Zero-shot learning and few-shot learning can be achieved by providing prompts to large language models. In the zero-shot setting, no specific instances are provided for the given task. Instead, the model makes predictions based solely on the provided instructional prompts and input statements. In contrast, in the K-shot setting, K instances for each class along with their labels are included in the input prompt. This approach enables the model to gain a better understanding of the task's intention and the desired answer format. We evaluated the effectiveness of transfer learning on two distinct datasets: *FCTR500*, which is more balanced, and *FCTR1000*, which is imbalanced. Note that in the experiments, we employed the models that were fine-tuned on the Snopes dataset with the corresponding results provided in Table 3.

Moreover, we conducted transfer learning experiments by repeating few-shot settings five times and reported the average scores along with the standard errors. According to Table 7 and Table 8, few-shot learning appears to be beneficial for the LLaMA variants. In other words, providing sample instances within prompts slightly enhanced their performance. However, fine-tuning LLaMA language models with Turkish data resulted in a substantial improvement in the F1-macro score. For instance, on the *FCTR1000* dataset, while few-

| Input | Model | F1-macro | F1-binary |
|---|---|---|---|
| zero shot | mBERT | 0.550 | 0.667 |
| zero shot | LLaMA-7B | 0.488 ∓ 0.026 | 0.577 ∓ 0.027 |
| 1-shot | LLaMA-7B | 0.536 ∓ 0.006 | 0.742 ∓ 0.009 |
| 2-shot | LLaMA-7B | 0.545 ∓ 0.035 | 0.632 ∓ 0.045 |
| 3-shot | LLaMA-7B | 0.577 ∓ 0.011 | 0.642 ∓ 0.029 |
| 4-shot | LLaMA-7B | 0.538 ∓ 0.021 | 0.609 ∓ 0.024 |
| 5-shot | LLaMA-7B | 0.533 ∓ 0.021 | 0.647 ∓ 0.022 |
| zero shot | LLaMA-13B | 0.498 ∓ 0.014 | 0.699 ∓ 0.006 |
| 1-shot | LLaMA-13B | 0.489 ∓ 0.026 | 0.683 ∓ 0.023 |
| 2-shot | LLaMA-13B | 0.530 ∓ 0.028 | 0.689 ∓ 0.019 |
| 3-shot | LLaMA-13B | 0.482 ∓ 0.022 | 0.670 ∓ 0.028 |
| 4-shot | LLaMA-13B | 0.529 ∓ 0.036 | 0.638 ∓ 0.028 |
| 5-shot | LLaMA-13B | 0.514 ∓ 0.013 | 0.632 ∓ 0.007 |
| zero shot | LLaMA-70B | 0.527 ∓ 0.042 | **0.773** ∓ 0.016 |
| 1-shot | LLaMA-70B | 0.507 ∓ 0.036 | 0.766 ∓ 0.018 |
| 2-shot | LLaMA-70B | 0.539 ∓ 0.021 | 0.754 ∓ 0.013 |
| 3-shot | LLaMA-70B | 0.492 ∓ 0.030 | 0.692 ∓ 0.023 |
| 4-shot | LLaMA-70B | 0.542 ∓ 0.021 | 0.709 ∓ 0.014 |
| 5-shot | LLaMA-70B | **0.585** ∓ 0.017 | 0.709 ∓ 0.023 |

Table 7: Transfer learning on the FCTR500 data

| Input | Model | F1-macro | F1-binary |
|---|---|---|---|
| zero shot | mBERT | 0.529 | 0.736 |
| zero shot | LLaMA-7B | 0.479 ∓ 0.019 | 0.647 ∓ 0.018 |
| 1-shot | LLaMA-7B | 0.501 ∓ 0.017 | 0.857 ∓ 0.013 |
| 2-shot | LLaMA-7B | 0.518 ∓ 0.010 | 0.706 ∓ 0.006 |
| 3-shot | LLaMA-7B | 0.501 ∓ 0.010 | 0.691 ∓ 0.024 |
| 4-shot | LLaMA-7B | 0.512 ∓ 0.023 | 0.694 ∓ 0.024 |
| 5-shot | LLaMA-7B | 0.502 ∓ 0.030 | 0.690 ∓ 0.048 |
| zero shot | LLaMA-13B | 0.502 ∓ 0.011 | 0.803 ∓ 0.006 |
| 1-shot | LLaMA-13B | 0.550 ∓ 0.016 | 0.811 ∓ 0.014 |
| 2-shot | LLaMA-13B | 0.539 ∓ 0.033 | 0.788 ∓ 0.020 |
| 3-shot | LLaMA-13B | 0.533 ∓ 0.017 | 0.763 ∓ 0.016 |
| 4-shot | LLaMA-13B | 0.537 ∓ 0.010 | 0.758 ∓ 0.010 |
| 5-shot | LLaMA-13B | 0.533 ∓ 0.029 | 0.737 ∓ 0.021 |
| zero shot | LLaMA-70B | 0.521 ∓ 0.018 | **0.865** ∓ 0.002 |
| 1-shot | LLaMA-70B | 0.528 ∓ 0.011 | 0.858 ∓ 0.011 |
| 2-shot | LLaMA-70B | **0.560** ∓ 0.033 | 0.841 ∓ 0.012 |
| 3-shot | LLaMA-70B | 0.536 ∓ 0.023 | 0.806 ∓ 0.018 |
| 4-shot | LLaMA-70B | 0.520 ∓ 0.019 | 0.808 ∓ 0.016 |
| 5-shot | LLaMA-70B | 0.521 ∓ 0.018 | 0.778 ∓ 0.015 |

Table 8: Transfer learning on the FCTR1000 data

shot learning achieved the highest F1-macro score of 0.560 (in Table 8), fine-tuning with Turkish data boosted all LLaMA variants to F1-macro score of 0.642 (in Table 5).

### 5.5. Neural Machine Translation

Neural machine translation is an approach that employs deep learning models to translate a text from a source language to a target language (Ranathunga et al., 2023). The transformer-based generative large language models are pretrained massively in English. Therefore, their performance in other languages may not be equally impressive. To tackle this challenge, we conducted translations of the Turkish fact-checking dataset into English utilizing the ChatGPT API. Table 9 presents the veracity detection results on the translated data. Note that we employed the models fine-tuned on the Snopes dataset.

| Dataset | Model | F1-macro | F1-binary |
|---|---|---|---|
| fctr500 | mBERT | 0.561 | **0.789** |
| fctr500 | LLaMA-7B | **0.576** ∓ 0.014 | 0.782 ∓ 0.007 |
| fctr500 | LLaMA-13B | 0.567 ∓ 0.018 | 0.739 ∓ 0.013 |
| fctr500 | LLaMA-70B | 0.571 ∓ 0.015 | 0.771 ∓ 0.007 |
| fctr1000 | mBERT | 0.485 | 0.840 |
| fctr1000 | LLaMA-7B | 0.524 ∓ 0.011 | 0.847 ∓ 0.003 |
| fctr1000 | LLaMA-13B | 0.573 ∓ 0.013 | 0.879 ∓ 0.004 |
| fctr1000 | LLaMA-70B | **0.581** ∓ 0.012 | **0.883** ∓ 0.003 |

Table 9: Turkish to English machine translation results

The results suggest that employing translated claims led to higher success rates for LLaMA models compared to the few-shot prompting approach. However, the success rate of mBERT was not positively influenced by translation. This phenomenon may be attributed to the differences in pretraining data between LLaMA models and mBERT. To be more specific, the LLaMA models were massively trained on English corpora, while the pretrained data for mBERT might exhibit a more uniform language distribution.

Additionally, we annotated the test set of *FCTR500* data based on claim statements, marking them as either ”local” or ”global”. Claims that specifically related to Turkiye were marked as ”local” claims, while claims with broader implications were labeled as ”global”. This categorization was done to assess the impact of the LLaMA model's pretrained knowledge on the claim category. We expected that the model would perform better on global claims, given the possibility that it might have pretrained information related to such claims from the web. The results indicate that using the LLaMA-13B model, the average F1-macro for local claims was 0.520 ∓ 0.036 while the average F1-macro score for global claims was 0.582 ∓ 0.056. However, using the LLaMA-7B model, we obtained the average F1-macro scores of 0.567 ∓ 0.017 for local claims and 0.541 ∓ 0.015 for global claims. The results imply that the higher F1-macro score for global claims with the larger LLaMA model may be attributed to its pretraining knowledge that should be addressed in further research.

Furthermore, we employed Opus-MT's (Tiedemann and Thottingal, 2020) *opus-mt-tc-big-en-tr* model to translate the Snopes dataset into Turkish and subsequently fine-tuned the language models using the translated Snopes' claims. This experiment was conducted to examine the impact of translating an English dataset into a low-resource

| Dataset | Model | F1-macro | F1-binary |
|---|---|---|---|
| fctr500 | mBERT | 0.532 | **0.757** |
| fctr500 | LLaMA-7B | 0.523 ∓ 0.019 | 0.630 ∓ 0.023 |
| fctr500 | LLaMA-13B | 0.544 ∓ 0.018 | 0.708 ∓ 0.006 |
| fctr500 | LLaMA-70B | **0.553** ∓ 0.025 | 0.725 ∓ 0.022 |
| fctr1000 | mBERT | 0.474 | 0.826 |
| fctr1000 | LLaMA-7B | 0.481 ∓ 0.023 | 0.705 ∓ 0.020 |
| fctr1000 | LLaMA-13B | 0.552 ∓ 0.044 | 0.800 ∓ 0.024 |
| fctr1000 | LLaMA-70B | **0.556** ∓ 0.018 | **0.832** ∓ 0.011 |

Table 10: English to Turkish machine translation results

language, specifically Turkish, on model performance. The fine-tuned models were then evaluated on the test splits of *FCTR500* and *FCTR100* to maintain consistency with the other experiments. According to Table 10, the F1-macro scores slightly decreased compared to the results presented in Table 9 when translating to a low-resource language.

Fine-tuning on translated data involves certain considerations. To be more specific, despite the state-of-the-art machine translation models accurately translating content, it might not be always feasible to maintain all context after translation. Additionally, since the current language models have a better understanding of English, it is an expected outcome that they would exhibit better performance on data translated from Turkish to English. Likewise, the results suggested that collecting native data for low-resource languages (Turkish for this case) is still required to ensure the development of successful models.

## 6. Discussion

The main objective of this study is to test the possibility and the extent of making use of a large amount of fact-checking data and large language models that were heavily pretrained in English for fact-checking in other languages with much less labeled data, and much smaller pretraining data for large language models. We focus on Turkish as a low-resource language for this task. Although focusing on a single familiar language allows us to curate a better fact-checking corpus, and perform more meaningful error analysis, our approach is applicable to many languages. Results are likely to differ based on typological similarity of the languages in question, as well other factors like geographical proximity and cultural similarity of the communities that speak the language.

Our experiments demonstrate some small gains from the high-resource language in zero-shot and few-shot settings, where few-shot learning shows slight improvement over zero-shot. The results in Table 7 and Table 8 shows a small but consistent increase in F1-macro scores when a few examples are included. The benefit of more few-shot examples is unclear, however. The same is true for making use of machine translation from low-resource language to high-resource language. The test instances translated to English labeled by the models trained on English data clearly better than an uninformed system. Even a small amount of training data provides better results than zero- or few-shot approaches.

Another interesting outcome of our results is the success of small models that rely only on surface cues on the FCTR data. There are no obvious latent variables (e.g., authors, source websites) that can identify the veracity label of short claim texts. This means some relevant information is available on the surface features. However, the large language models surpass the simple ones on English with a large margin (see Table 3). This may indicate both the help of the linguistic and perhaps factual information brought by these models.[12] However, most probably the comparatively smaller Turkish data during pretraining is possibly a factor in low scores of LLaMA with fine-tuning with Turkish (Tables 4 and 5).

In the majority of the experiments, only the claim statements were employed as input, since this is a more realistic scenario as individuals typically seek to assess the truthfulness of a claim before spending time gathering additional information. We also include evidence statements as input in some experiments, which show a clear benefit in providing additional information. However, evidence retrieval is also a challenging problem in fact-checking (which falls beyond the scope of this study). A further problem with providing evidence may be discouraging the model from leveraging its pretrained knowledge while making decisions.

## 7. Conclusion

We present a novel Turkish fact-checking dataset that is collected from three fact-checking resources. It includes 3238 claims with additional metadata from the same resources including evidence and summary of the justifications. The experiments revealed that fine-tuning a large language model on the Turkish dataset yields superior results compared to the zero-shot and few-shot approaches, highlighting the importance of employing datasets for languages with limited resources.

[12] A potential problem here is these models may have the full fact-checking report for the test instances, including the clearly stated verdict in their pretraining data.

## 8. Ethical Considerations and Limitations

First, we did not process the collected data to ensure anonymization. The dataset encompasses fact-checked claims about public figures including politicians and artists. If any individual mentioned in a claim requests their removal, we can eliminate the associated claims.

Secondly, the data acquisition process adhered to the regulations of the Turkish text and data mining policy. This policy underlies that the datasets can be used exclusively for research purposes.

Lastly, the Snopes dataset was collected in accordance with the Terms of Use set by Snopes. Therefore, anyone interested in accessing the Snopes dataset must send a request that includes a commitment to use the dataset only for non-commercial purposes.

## 9. Acknowledgements

This research is supported by the Scientific and Technological Research Council of Turkey (TUBITAK, Prog: 2214-A) and the German Academic Exchange Service (DAAD, Prog: 57645447). We would like to thank the anonymous reviewers for their suggestions to improve the study. We also appreciate METU-ROMER and the University of Tübingen for providing the computational resources.

Parts of this research received the support of the EXA4MIND project, funded by the European Union´s Horizon Europe Research and Innovation Programme, under Grant Agreement N° 101092944. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

## 10. Bibliographical References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14940–14949.

Naser Ahmadi, Thi-Thuy-Duyen Truong, Le-Hong-Mai Dao, Stefano Ortona, and Paolo Papotti. 2020. RuleHub: A public corpus of rules for knowledge graphs. *Journal of Data and Information Quality (JDIQ)*, 12(4):1–22.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020a. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020b. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697.

Alberto Barrón-Cedeno, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. Task 2: Factuality. *CLEF (Working Notes)*, 2125.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Erik Brand, Kevin Roitero, Michael Soprano, Afshin Rahimi, and Gianluca Demartini. 2022. A neural model to jointly predict and explain truthfulness of statements. *ACM Journal of Data and Information Quality*, 15(1):1–19.

Recep Firat Cekinel and Pinar Karagoz. 2024. Explaining veracity predictions with evidence summarization: A multi-task model approach. *arXiv preprint arXiv:2402.06443*.

Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2021. Causal understanding of fake news dissemination on social media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 148–157.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 492–502.

Shih-Chieh Dai, Yi-Li Hsu, Aiping Xiong, and Lun-Wei Ku. 2022. Ask to know more: Generating counterfactual explanations for fake claims. *arXiv preprint arXiv:2206.04869*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.

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.

Jiangshu Du, Yingtong Dou, Congying Xia, Limeng Cui, Jing Ma, and S Yu Philip. 2021. Cross-lingual COVID-19 fake news detection. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 859–862. IEEE.

Mohamed H Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. Exfakt: A framework for explaining facts over knowledge graphs and text. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 87–95.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. Xhate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-Fact: A new benchmark dataset for multilingual 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 2: Short Papers)*, pages 675–682.

Samar Haider, Luca Luceri, Ashok Deb, Adam Badawy, Nanyun Peng, and Emilio Ferrara. 2023. Detecting social media manipulation in low-resource languages. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1358–1364.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503.

Mareike Hartmann, Yevgeniy Golovchenko, and Isabelle Augenstein. 2019. Mapping (dis-)information flow about the MH17 plane crash. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 45–55.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and S Yu Philip. 2022. Chef: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376.

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. Concrete: Improving cross-lingual fact-checking with cross-lingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3):598–608.

Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuiseok Lim. 2019. exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (BERT). *Applied Sciences*, 9(19):4062.

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.

Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, Scott A Hale, and Rada Mihalcea. 2022. Matching tweets with applicable fact-checks across languages. *arXiv preprint arXiv:2202.07094*.

Kashif Khan, Ruizhe Wang, and Pascal Poupart. 2022. WatClaimCheck: A new dataset for claim entailment and inference. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1304, Dublin, Ireland. Association for Computational Linguistics.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. FactKG: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206, Toronto, Canada. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.

Yichuan Li, Kyumin Lee, Nima Kordzadeh, and Ruocheng Guo. 2023. What boosts fake news dissemination on social media? a causal inference view. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 234–246. Springer.

Hongzhan Lin, Pengyao Yi, Jing Ma, Haiyun Jiang, Ziyang Luo, Shuming Shi, and Ruifang Liu. 2023. Zero-shot rumor detection with propagation structure via prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5213–5221.

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLIPpings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157.

Dan S Nielsen and Ryan McConville. 2022. MuMiN: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3141–3153.

Jeppe Nørregaard and Leon Derczynski. 2021. DanFEVER: claim verification dataset for Danish. In *Proceedings of the 23rd Nordic conference on computational linguistics (NoDaLiDa)*, pages 422–428.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2532–2544, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeff Z Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *International semantic web conference*, pages 669–683. Springer.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.

David Pogue. 2017. How to stamp out fake news. *Scientific American*, 316(2):24–24.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023. Fake news detection in Dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126:106877.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. (Mis)information dissemination in WhatsApp: Gathering, analyzing and countermeasures. In *The World Wide Web Conference*, pages 818–828.

Mourad Sarrouti, Asma Ben Abacha, Yassine M'rabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643.

Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid–a multilingual cross-domain fact check news dataset for COVID-19. *arXiv preprint arXiv:2006.11343*.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Propagation2vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management*, 58(5):102618.

Niraj Sitaula, Chilukuri K Mohan, Jennifer Grygiel, Xinyi Zhou, and Reza Zafarani. 2020. Credibility-based fake news detection. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 163–182. Springer.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32–43.

Lukas Stappen, Fabian Brunn, and Björn Schuller. 2020. Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and AXEL. *arXiv preprint arXiv:2004.13850*.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake news detectors are biased against texts generated by large language models. *arXiv preprint arXiv:2309.08674*.

S Suryavardan, Shreyash Mishra, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, et al. 2023. Factify 2: A multimodal fake news and satire news dataset. *arXiv preprint arXiv:2304.03897*.

Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. 2021. New explainability method for BERT-based model in fake news detection. *Scientific Reports*, 11(1):1–13.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2944–2953.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114.

Herbert Ullrich, Jan Drchal, Martin Rỳpar, Hana Vincourová, and Václav Moravec. 2023. Csfever and ctkfacts: acquiring czech data for fact verification. *Language Resources and Evaluation*, pages 1–35.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304.

William Yang Wang. 2017. “Liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2023. Counterfactual debiasing for fact verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6777–6789, Toronto, Canada. Association for Computational Linguistics.

Jing Yang, Didier Vega-Oliveros, Taís Seibt, and Anderson Rocha. 2022. Explainable fact-checking through question answering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8952–8956. IEEE.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Weifeng Zhang, Ting Zhong, Ce Li, Kunpeng Zhang, and Fan Zhou. 2022. Causalrd: A causal view of rumor detection via eliminating popularity and conformity biases. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1369–1378. IEEE.

Wenxuan Zhang, Yang Deng, Jing Ma, and Wai Lam. 2020. Answerfact: Fact checking in product question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2407–2417.

Xinyi Zhou, Atishay Jain, Vir V Phoha, and Reza Zafarani. 2020. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2):1–25.

Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD explorations newsletter*, 21(2):48–60.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

# A. Topic Modeling

| Dataset | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 |
|---|---|---|---|---|---|---|
| FCTR500-train | 39 | 64 | 105 | 49 | 116 | 27 |
| FCTR500-val | 8 | 10 | 10 | 9 | 9 | 4 |
| FCTR500-test | 6 | 9 | 7 | 9 | 15 | 4 |
| FCTR1000-train | 73 | 132 | 174 | 130 | 237 | 54 |
| FCTR1000-val | 9 | 16 | 20 | 18 | 29 | 8 |
| FCTR1000-test | 12 | 11 | 19 | 21 | 35 | 2 |
| FCTR | 293 | 472 | 524 | 600 | 927 | 167 |

Table 11: Topic distribution in the FCTR dataset

| Dataset | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 |
|---|---|---|---|---|---|---|---|
| Snopes-train | 206 | 1063 | 386 | 260 | 553 | 327 | 193 |
| Snopes-val | 26 | 125 | 52 | 27 | 73 | 48 | 23 |
| Snopes-test | 25 | 124 | 43 | 29 | 75 | 50 | 27 |

Table 12: Topic distribution in the Snopes dataset

| Topics | Representative Words (transl.) |
|---|---|
| Topic1 | claim, news, person, sharing, information, account, share, be, child, use |
| Topic2 | photograph, image, account, sharing, share, claim, video, name, view, use |
| Topic3 | country, Turkiye, year, history, claim, data, take, be, state, Turkic |
| Topic4 | vaccine, be, virus, claim, work, human, disease, research, person, impact |
| Topic5 | video, claim, news, be, statement, sharing, name, history, eat, talk |
| Topic6 | use, product, breeding, water, electricity, plane, production, year, logo, claim |

Table 13: Representative words in *FCTR* dataset

| Topics | Representative Words |
|---|---|
| Topic1 | animal, water, world, report, military, human, fire, Russian, area, Russia |
| Topic2 | say, people, year, man, know, take, make, time, go, get |
| Topic3 | image, photograph, show, video, picture, take, create, appear, film, real |
| Topic4 | Trump, president, Obama, White House, former, Clinton, President Donald, tweet, Donald Trump, say |
| Topic5 | post, article, news, Facebook, claim, story, publish, report, page, com |
| Topic6 | state, law, government, report, vote, bill, United States, federal, election, claim |
| Topic7 | covid, vaccine, health, study, drug, medical, cause, disease, use, patient |

Table 14: Representative words in *Snopes* dataset

Topic modeling is a method for discovering abstract topics in a collection of documents. Latent topics indicate the patterns in the data that can be inferred by the relationships between words that occur in the documents. The output of a topic modeling is a set of abstract topics that are represented by a list of the most representative words in the topic. In our analysis, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic modeling is applied to the *Snopes* and *FCTR* datasets to explore the latent patterns using the coherence metric. The coherence score can be used to evaluate the semantic similarity between the words in a topic.

The topic distributions for each data split are given in Table 11 and Table 12 respectively. Even though we did not split the datasets according to the topic ratios, the most dominant and the least frequent topics were preserved in all data splits. For instance, in the *FCTR* dataset, The fifth topic is the most frequent topic in all subsets except *FCTR500-val* in which the given topic is not the most dominant topic by a small margin. Additionally, the sixth topic is the least frequent topic in all splits.

We utilized lemmatization, employing the Spacy library for English [13] and the Zeyrek library for Turk-

[13] https://spacy.io/models/en

| Subset | Feature name | Adjusted p-value |
|---|---|---|
| FCTR500 | allcaps | 0.023 |
| FCTR500 | avg_wordlen | 0.018 |
| FCTR500 | coleman_liau_index | 0.018 |
| FCTR500 | lix | 0.032 |
| FCTR1000 | NNP | 0.049 |
| FCTR1000 | avg_wordlen | 0.048 |
| FCTR1000 | coleman_liau_index | 0.045 |
| FCTR1000 | lix | 0.048 |

Table 15: Statistically significantly different NELA features

ish [14]. Table 13 and Table 14 display the most representative words for each topic. The coherence score for the Turkish dataset within these topics was 0.388, and the perplexity score was -7.699. The average entropy value per document was calculated as 1.50, suggesting a moderate topic distribution level. Similarly, the Snopes dataset achieved a coherence score of 0.450 and a perplexity score of -8.796. Moreover, the average entropy score per document was found to be 1.94 which might indicate that the documents cover multiple related topics without a strong focus on a single one.

## B. NELA Features

News Landscape (NELA) features (Horne and Adali, 2017) are manually crafted content-based textual attributes for news veracity detection. The authors divided the features into six classes: style, complexity, bias, affect, moral and event. We applied NELA features to examine the discrepancies of the features for fake and true claims in the FCTR dataset and conducted Tukey's pairwise test (Tukey, 1949) to identify statistically significant differences.

Table 15 presents features that exhibit statistically significant distinctions for *FCTR500* and *FCTR1000*. We computed the NELA features for only claim statements and the results indicate that only a few features demonstrate significant divergence for fake and true claims.

[14] https://zeyrek.readthedocs.io/en/latest/