ViClaim: A Multilingual Multilabel Dataset for Automatic Claim Detection in Videos

Patrick Giedemann[1], Pius von Däniken[1], Jan Deriu[1],
Alvaro Rodrigo[2], Anselmo Peñas[2], Mark Cieliebak[1]
[1]Zurich University of Applied Sciences, Winterthur
[2] UNED NLP & IR Group, Spain
gied@zhaw.ch

Abstract

The growing influence of video content as a medium for communication and misinformation underscores the urgent need for effective tools to analyze claims in multilingual and multi-topic settings. Existing efforts in misinformation detection largely focus on written text, leaving a significant gap in addressing the complexity of spoken text in video transcripts. We introduce ViClaim, a dataset of 1,798 annotated video transcripts across three languages (English, German, Spanish) and six topics. Each sentence in the transcripts is labeled with three claim-related categories: fact-check-worthy, fact-non-check-worthy, or opinion. We developed a custom annotation tool to facilitate the highly complex annotation process. Experiments with state-of-the-art multilingual language models demonstrate strong performance in cross-validation (macro F1 up to 0.896) but reveal challenges in generalization to unseen topics, particularly for distinct domains. Our findings highlight the complexity of claim detection in video transcripts. ViClaim offers a robust foundation for advancing misinformation detection in video-based communication, addressing a critical gap in multimodal analysis.

1 Introduction

Video content is increasing in popularity worldwide. Alone in the US, the average adult spends around 47 minutes on YouTube and 55 Minutes on TikTok [1]. Especially short-format videos (i.e., videos of at most 90 seconds) are gaining in popularity [2]. While platforms like YouTube serve educational and informational purposes (Srinivasacharlu, 2020; Wafa'A and Khasawneh, 2024), they are also used to disseminate narratives aimed at influencing viewer's opinions. Despite the recognized necessity of extending misinformation detection technologies to video modalities (Da San Martino et al., 2021), research efforts remain limited, with a predominant focus on analyzing textual data from platforms like X (formerly Twitter) (Arslan et al., 2020; Alam et al., 2023). Furthermore, most existing research on video-based misinformation primarily targets visual features, such as detecting manipulated content (Papadopoulou et al., 2018; Palod et al., 2019; Bu et al., 2023), leaving the semantic analysis of video transcripts largely unexplored, although features found in transcripts of spoken language differ from written language in structure and delivery (Dingemanse and Liesenfeld, 2022). Studies that do analyze transcripts tend to treat them as a global, binary classification task, aiming to determine whether a video contains misinformation as a whole (Hou et al., 2019; Hussein et al., 2020; Papadamou et al., 2022; Christodoulou et al., 2023). However, misinformation detection is a multilayered task, with claim detection as the first step (Panchendrarajan and Zubiaga, 2024).

In this work, we present the first step towards misinformation detection in videos by introducing ViClaim, a novel dataset where transcripts of YouTube Short videos are annotated using a custom annotation tool on a sentence level with the claim taxonomy introduced in (Panchendrarajan and Zubiaga, 2024). Each sentence is annotated to indicate whether it contains a claim, an opinion, or both, and if a claim is present, whether it is check-worthy. Since sentences may fall into multiple categories simultaneously, the task is framed as a multi-label classification problem. The dataset encompasses 1,798 videos in three languages (English, German, and Spanish) across six topics—five political and one entertainment—allowing for investigating domain transfer capabilities. Each sentence was annotated by four independent annotators, resulting in a total of 17,116 annotated sentences. Thus, ViClaim offers a rich database for future research

---

[1]https://www.statista.com/statistics/1359403/us-time-spent-per-day-netflix-tiktok-youtube
[2]https://www.yaguara.co/short-form-video-statistics/

on multimodal misinformation detection. As a first step to showcase the utility of ViClaim, we trained several baseline models on the transcripts (leaving multimodal models for future work). Our best-performing model achieved an F1 score of 89.9 for checkworthiness classification, 77.6 for detecting non-check-worthy claims, and 83.6 for opinion detection. These results demonstrate both the effectiveness of the dataset and the challenges posed by the nuanced task of analyzing short-form video transcripts. To summarize, our contributions are as follows: (i) the ViClaim dataset, a multilingual, multi-topic resource for claim detection in video transcripts; (ii) a custom annotation tool and comprehensive guidelines; (iii) baseline models to showcase the potential of ViClaim.

Release. We release the corpus in form of the video IDs used, the time stamps of the annotated sentences, the corresponding labels, and the code pipeline to reconstruct the corpus, which can be found under the following GitHub repository [3]. We release the experimental pipeline and the trained models, which can be found under the following GitHub repository [4].

2 Related Work

Claim detection constitutes the first step in misinformation detection (Panchendrarajan and Zubiaga, 2024) where we mainly refer to claims that, due to their content, have a significant impact on public opinion or pose a risk of being widely disseminated due to their controversial nature (Alam et al., 2023). There has been vast research on claim detection, mainly on texts (Zhang and Ghorbani, 2020), where most systems receive the exact statement to be checked. These claims are usually sentences extracted from a document or social media posts (e.g., a Tweet) (Barrón-Cedeno et al., 2020). On the other hand, other datasets aimed at detecting the exact span containing the claim in tweets, which is a more challenging task, but it allows systems to retrieve more accurate evidence (Sundriyal et al., 2022). The available datasets commonly focused on a single and controversial topic like environment (Stammbach et al., 2023), politics (Dutta et al., 2023) or COVID-19 (Salek Faramarzi et al., 2023). Some datasets contain claims about several topics like elections or COVID-19 (Kazemi et al., 2021). The most common language for these datasets is

English. Still, there have also been several efforts for creating datasets in multiple languages, like the dataset from the NLP4IF 2021 shared task, with tweets in English, Arabic, and Bulgarian (Shaar et al., 2021), extended in the CLEF2022 Check-That! lab with Dutch tweets (Alam et al., 2021).

Although most of the work on claim detection has been devoted to text content, there have also been some efforts towards detecting claims in multimedia content. These works consider that multimedia content in social networks is an effective way for spreading misinformation compared with textual content (Jin et al., 2017; Dhawan et al., 2022). One of the first works in the multimedia setting extended previous text-based collections with images to evaluate multimodal detection approaches (Cheema et al., 2021). However, the labels were only based on the textual content. This is why Cheema et al. (2022) created the MM-Claims dataset, which includes tweets and corresponding images and where the annotations are based on both text and images. The CheckThat! Lab 23 followed their methodology (Alam et al., 2023), where the organizers show the tweet's image alongside the tweet itself, and the image could be a piece of evidence or contain a text containing a claim. The inclusion of multimodal data yielded better scores (von Däniken et al., 2023).

On the other hand, to the best of our knowledge, there is almost no work on claim detection using videos. The most relevant work is the ClaimBuster dataset (Arslan et al., 2020), which provides 23,533 statements extracted from transcripts of US general election presidential debates and annotated by humans for check-worthiness. This work is somewhat similar to our dataset, although it only provides the statements to be checked without context, in English, and without the original audio or video. Most of the previous works related to videos have focused on the task of detecting whether they contain misinformation (Hou et al., 2019; Hussein et al., 2020; Papadamou et al., 2022; Christodoulou et al., 2023). Micallef et al. (2022) found that traditional approaches focused on textual content and missed post-video pairs that contain misinformation. They improved detection results when considering features from platforms containing the linked videos. Choi and Ko (2022) developed a deep learning model that integrates different modalities for detecting if full YouTube videos were fake or real. Palod et al. (2019) developed VAVD, a video dataset with fake and non-fake annotations. Besides, they ob-

---

[3]`https://github.com/pgied/viclaim_stt`
[4]`https://github.com/pgied/viclaim_training`

tained promising results by relying only on user comments. Papadopoulou et al. (2018) created the InVID Fake Video Corpus, with annotations for the full videos.

3 Data Collection

This section describes the data collection efforts for claim detection in video. The final corpus contains 1798 videos in short format. Each video was transcribed and split up into sentences, which resulted in a total of 17116 sentences. Each sentence has been annotated by four different annotators. The dataset spans three languages: English, German, and Spanish [5]. We first describe the annotation tool together with the task; then, we describe the label set, the topics of the videos, the video selection, and the annotator agreement computation. The data collection ran from 12. April to 3. June 2024 and from 28. October to 25. November 2024.

3.1 Annotation Tool

The annotation tool (see Figure 1) displays the video alongside its transcript. Annotators are tasked with watching the video and labeling each sentence using the provided tagset: fact-check-worthy, fact-non-check-worthy, opinion, or none (cf. 3.2). Following the approach of (Arslan et al., 2020), we collect annotations at the sentence level, as allowing annotators to define their own spans often introduces excessive noise.

While we recognize that sentences may contain multiple claims (some check-worthy, some not) or combine opinions and claims and that some may span multiple sentences (see Table 1), we address these complexities by employing a multi-label annotation approach. This ensures comprehensive coverage without compromising consistency. Importantly, we tested allowing annotators to define custom spans, which led to significant disagreement and unusable data due to varying interpretations of span boundaries and overlaps. As a result, sentence-level annotation was chosen as a more consistent and reliable alternative [6].

The transcripts were created using AssemblyAI [7], and the sentence segmentation was per-

formed using SpaCy [8].

3.2 Tag-Set Description

Our tagset is based on the taxonomy introduced in (Panchendrarajan and Zubiaga, 2024), which categorizes claims into factual and opinions, and factual are further divided into check-worthy facts and non-check-worthy (either due to being non-verifiable or not check-worthy). Thus, we introduce the three labels: Fact Check-worthy (FCW), Fact Non-Check-worthy (FNC), and Opinion (OPN). Note that in contrast to (Arslan et al., 2020), which only differentiates between Check-worthy Factual Statements and Uncheckable Factual Statements, we explicitly label opinions to capture subjective elements as well as broader contextual and persuasive aspects that are essential for understanding misinformation (Goldberg and Marquart, 2021). Additionally, we adopt a multi-label annotation approach, recognizing that sentences often encompass multiple relevant categories.

Fact that is check-worthy (FCW). Sentences containing factual claims of public interest that are verifiable and relevant for fact-checking, commonly sought by journalists.

Fact that is not check-worthy (FNC). Factual claims that are either unverifiable or lack public interest, such as personal experiences or jokes.

Opinion (OPN). This tag encompasses subjective sentences, including opinions, beliefs, accusations, speculations, predictions, and emotional expressions.

None. Sentences that do not fit the above categories, such as commands, insults, casual expressions, or threats.

For examples of each tag and their combinations, see Table 1.

3.3 Topic Selection

We selected five highly relevant socio-political topics during the video selection process conducted in May 2024 and November 2024. These topics were chosen based on their international relevance and widespread public interest, although it is important to acknowledge that the selected videos reflect a bias toward the Western world. The aim was to ensure coverage of topics that would generate content in multiple languages, enabling a diverse and multilingual dataset. Below, we describe the selected topics:

---

[5] The language selection is based on the languages spoken by the authors of this work, which allowed us to monitor the annotation process.

[6] Previous experiments with custom spans resulted in Krippendorff's α scores below 0.2, making the approach impractical.

[7] https://www.assemblyai.com. We opted for a paid solution as it provides diarization out of the box.
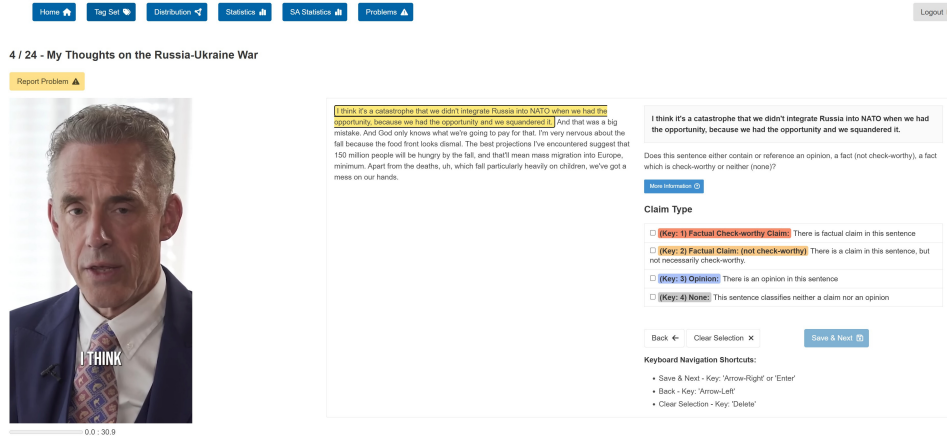
[8] https://spacy.io

Figure 1: Annotation Tool. The user is shown a short video and the transcript, which is split by sentence, and then they annotate each sentence with one or more of the four tags.

Sentence	Labels	Explanation
10,000 immigrants arrive daily.	FCW	Clearly checkable number and relevant to society.
And that's what people over and over told me, that, of course.	FNC	This claim cannot be verified and is therefor not check-worthy.
The company will probably go bankrupt within a year.	OPN	This is clearly an opinionated speculation.
Don't ever use the word smart with me.	NONE	This is a threat and is neither an opinion nor a fact.
He's first to admit that, uh, and he's pretty profane at times when he's fired up about something, and certainly he is about Donald Trump.	FNC, OPN	This sentence contains the labels FNC and OPN, since the first part is not possible to check, there is not verifiable information and the middle part is a subjective view (opinion).
These politicians will lie to your face and make millions while normal Americans pay the price.	FCW, OPN	This sentence contains the labels FCW and OPN, since the first part is a subjective view and the second part is verifiable and has a public relevancy.
I don't have them in front of me, but we're going to, if Pence becomes a candidate, we will look at that in more detail.	FNC, OPN	This sentence contains the labels FNC and OPN, since the first part is a self description and not not really publicly relevant and the second part is a future prediction which is not yet a fact but rather an intention.

Table 1: Examples of sentences within the transcripts along with their ground-truth labels, and explanations

- US Elections 2024. This topic includes videos focusing mostly on the two leading candidates, Donald Trump and Joseph Biden [9]. The focus on these candidates allows recognition beyond the United States, as we included videos in both German and Spanish in addition to English.
- War in Ukraine. This topic covers the conflict between Ukraine and Russia and has generated videos in multiple languages, making it suitable for our multilingual dataset.
- Migration. Videos on this topic discuss various aspects of migration, providing content in multiple target languages.
- European Union. This topic includes perspectives on the EU, particularly around the European Parliament elections in June 2024, with videos available in English, German, and Spanish.
- General view about the USA. This category includes perspectives on US society from both internal and external viewpoints, addressing sociopolitical dynamics.

To evaluate the claim detection approach in a domain transfer setting, we included an unrelated topic: League of Legends [10], a globally popular online multiplayer video game. This topic was selected to evaluate the dataset's applicability to different domains and support experiments in out-of-domain generalization.

3.4 Video Selection

Creating video annotations requires significant effort and resources, as it depends on identifying

---

[9]Since most of the video selection was conducted in May 2024, not many videos account for the candidacy of Kamala Harris.

[10]https://www.leagueoflegends.com/de-de/

videos that explicitly contain claims. Preliminary tests revealed that semi-automated approaches, such as keyword-based searches, often returned videos irrelevant to the target topics or devoid of claims. We opted for a manual video selection process due to the inherent challenges of using YouTube's recommender system, as highlighted by (Chandio et al., 2024) and our experience. Their findings show that factors like strong recency bias, the choice of seed videos, and the depth of exploration significantly influence the diversity and characteristics of recommended content. These complexities make automated selection prone to biases that are difficult to control, potentially limiting its effectiveness in curating a representative dataset.

To address these challenges, two trained researchers fluent in all three target languages manually searched for short-form videos (no longer than 90 seconds) across six predefined topics. New YouTube accounts were created to minimize recency and personalization biases. They used structured keyword searches tailored to each topic and refined keywords iteratively to ensure diversity. For example, keywords for the US Elections included "US elections 2024 Biden," "Trump policies," and "election debates." A mix of popular and niche videos was selected to capture diverse content and viewpoints, particularly for contentious topics like Migration and the US Elections.

The manual selection yielded 1798 short-form videos across six topics and three languages (approx. 100 videos per topic-language pair, cf. Appendix B).

3.5 Annotation Management

Due to the complexity of the annotation task, we opted against crowdsourcing and contracted twelve annotators (6 female, 6 male, ages 18–34). Our annotators included seven native German speakers (two also proficient in Spanish), two native Spanish speakers, and three bilingual English/German or Spanish/German speakers. All but two were students, and all were highly proficient in English, German, or Spanish. Each annotator was assigned 600 videos, grouped as follows:

- Group 1: Four annotators annotated 300 videos in English and 300 in German.
- Group 2: Four annotators annotated 300 videos in English and 300 in German (distinct from Group 1).

- Group 3: Four annotators annotated 600 videos in Spanish.

We compensated the annotators with 500 euros for a clean completion. This corresponds to an hourly salary of 25 Euros. Following Bender and Friedman (2018), we let the annotators complete a questionnaire to rate our task and collect information about their stances on the six topics (a properly anonymized version is available in Appendix F).

Participant Training. The training process contained guidelines, examples, workshops, and ongoing support for annotators throughout the annotation process.

First, we provided the annotators with guidelines to instruct them on the usage of the annotation tool, the tagset explanation (cf. 3.2), and examples, including edge cases and ambiguous examples. The set of examples was dynamically extended when new edge cases were discovered. We began the training of the annotators with a kickoff workshop to introduce annotators to the project, annotation tool, and task requirements. Annotators were asked to complete 10–20 annotations within the first two days. This allowed us to monitor their understanding and provide targeted feedback based on their early performance. The second workshop occurred after the first 100–200 annotations. Annotations with low agreement were reviewed and discussed in detail to identify patterns of misunderstanding. We maintained an open line of communication, promptly addressing any emerging questions or challenges. This proactive approach ensured annotators remained confident and consistent in their work.

3.6 Quality Control

Since the task is inherently ambiguous, we established the following process to ensure annotation quality. Similar to (Arslan et al., 2020), two authors collaboratively created a gold standard by annotating 30 videos per language (5 per topic), which resulted in 833 gold sentences. This collaborative effort ensured a consistent interpretation of claims across topics and languages. We continuously tracked annotator agreement with the gold standard, intervening directly (by reaching out to the annotators) when deviations occurred to maintain alignment and improve annotator understanding.

To further ensure the reliability of the annotations, we monitored inter-annotator agreement

(Krippendorff's α) and pairwise agreement using Jaccard similarity scores for the multi-label annotations over time. This continuous tracking allowed us to detect and address inconsistencies early in the process, thereby improving the overall quality of the dataset.

Annotator Agreement. We evaluated annotator agreement using two metrics. First, we calculated Cohen's κ between individual annotators and the gold standard annotations. Second, we computed Krippendorff's α to measure agreement across all annotators. Table 2 presents the agreement scores for each of the three groups.

For Cohen's κ, we report the average and maximum values across the four annotators in each group. Agreement levels varied by label:

- FCW: Agreement is moderate on average $(0.4 - 0.59)$, with Group 3 achieving substantial agreement $(0.64 - 0.76)$.
- FNC: Agreement is lower, with fair levels on average $(0.35 - 0.52)$, though Group 3 achieved a maximum of 0.69.
- OPN: Agreement is moderate $(0.47 - 0.58$ on average), with Group 3 again demonstrating higher consistency $(0.58 - 0.66)$.
- None: This label consistently achieved the highest agreement, ranging from $0.57 - 0.85$, with averages close to substantial levels.

For Krippendorff's α, scores ranged from 0.415 (Group 2) to 0.522 (Group 3), reflecting moderate agreement overall. Although these agreement levels are not particularly high, they are consistent with the inherent difficulty and subjectivity of the task. A closer analysis revealed that most ambiguity stems from the scenario where multiple classes are appropriate, and two annotators chose non-overlapping subsets of these appropriate classes (see Appendix C for examples of such ambiguous sentences). Thus, the disagreement often did not stem from poor annotator behavior but from the inherent ambiguity. Generally, it has been shown that differentiating between factual statements and opinions is a difficult task (Goldberg and Marquart, 2021).

In our case, at least one annotator in each group consistently demonstrated moderate to substantial agreement with the gold standard annotations, which we used as a basis for label normalization. This ensures that the resulting dataset retains high

	Group 1	Group 2	Group 3
FCW	0.51 | 0.59	0.57 | 0.70	0.64 | 0.76
FNC	0.38 | 0.52	0.35 | 0.48	0.42 | 0.69
OPN	0.47 | 0.53	0.53 | 0.68	0.58 | 0.66
None	0.70 | 0.77	0.57 | 0.85	0.68 | 0.81
α	0.442	0.415	0.522

Table 2: Agreement scores. Cohen's κ between the annotators and the gold annotations, we report the average and maximum overall annotators for each group. The last column shows the Krippendorf α-score for the annotator agreement.

reliability despite the task's unavoidable ambiguity challenge.

Label Normalization. To generate the final labels for the dataset, we leverage the observation that some annotators demonstrate higher agreement with the gold standard annotations. To account for this, we use MACE (Hovy et al., 2013), a Bayesian Model used to compute a trust score for each annotator, reflecting their reliability. We normalize these trust scores to sum to 1 across all annotators. **Soft Label.** Based on the trust scores, we derived soft-labels, which capture the ambiguity of the annotations. The soft label, denoted as $y_{\text{soft}} \in [0, 1]^3$, provides a probabilistic interpretation of the label assignments. Each entry represents the probability of the corresponding label. To create the soft label, we simply compute $y_{\text{soft}} = \sum_i w_i * a_i$, where w_i denotes the normalized trust score of annotator i and $a_i \in \{0, 1\}^3$ is the annotation of annotator i. This approach ensures that the soft label incorporates annotator trust and provides nuanced information about the likelihood of each label.

3.7 Data Overview

The resulting dataset contains 1,798 videos, approximately 100 videos per language and topic. In total, there are 17,116 annotated sentences. Figure 2 illustrates the distribution of labels across the six topics.

While political topics, such as the US Elections and the War in Ukraine, have many Fact-check-worthy (FCW) samples, the League of Legends topic exhibits a significantly lower prevalence of FCW labels. Instead, it shows a higher prevalence of Fact-Non-check-worthy (FNC) labels. This variation highlights the differences like the topics, with political content being more focused on verifiable claims, while non-political content tends to include
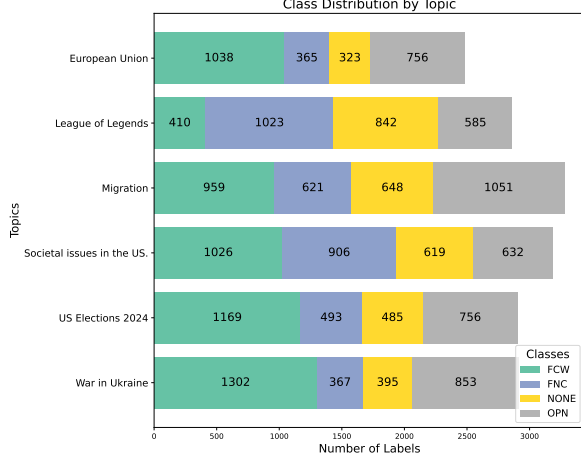
Figure 2: For each label, each topic's appearance percentage is depicted. The labels are sorted by their overall frequency.

more subjective or non-verifiable statements.

4 Baseline Experiments

Here, we describe the baseline experiments, for which we fine-tune four different pre-trained multilingual large-language models (LLM). These experiments serve as starting points, leaving more sophisticated and multimodal approaches to future work.

4.1 Experimental Setting

Data. The models are trained to predict the labels of each sentence in a transcript. For this, let $\mathcal{C} = \{c_i\}_{i=1}^{1798}$ be the set of all clip transcripts. Each transcript is segmented into sentences that are labeled, thus $c_i = \{s_j^{(i)}\}_{j=1}^{n_i}$ where n_i is the number of sentences of transcript c_i. Each sentence is a pair of text and label $s_j^{(i)} = (x_j^{(i)}, y_j^{(i)})$, where $x_j^{(i)} = \{t_k\}_{k=1}^{m_{ij}}$ denotes the tokens (m_{ij} number of tokens in sentence j of clip i), and $y_j^{(i)}$ is the label (cf. 3.6). The input to the classifiers is the concatenation of the full clip transcript for the context, with the sentence of the transcript to be classified, i.e., $\mathcal{D} = \{([x_1^{(i)}; ..; x_{n_i}^{(i)}; x_j^{(i)}], y_j^i)\}$, where i denotes the clip number, j^{th} the sentence within clip i to be classified, and $[;]$ denotes the concatenation of sentence strings. During training, we use the soft labels as in (Fornaciari et al., 2021) by computing the cross-entropy loss between the system output and the soft labels.

Models. We selected four state-of-the-art LLMs to be fine-tuned; they are all pre-trained on multilingual data containing our three languages of interest.

- **XLM-Robertal-Large (XLM).** Conneau et al. (2020) pre-trained an 550M parameter encoder-transformer on 100 languages.
- **Falcon-7B (F7B).** Almazrouei et al. (2023) pre-trained a 7B decoder-only model on 1.5T tokens of the RefinedWeb corpus (Penedo et al., 2023). The main languages that Falcon performs well on are English, German, Spanish, and French, which cover our use case well.
- **Mistral-7B (M7B).** Jiang et al. (2023) pre-trained a 7B parameter decoder-only model. The details of the training data used are undisclosed.
- **LLama3.2-3B (L3B).** Grattafiori et al. (2024) pre-trained a 3B parameter decoder-only model trained on an undisclosed set of 15T tokens of web data.

While we apply regular fine-tuning on *XLM-Roberta-Large*, we use Quantization and Low-Rank Adapters (QLoRA) (Dettmers et al., 2024) to fine-tune the three LLMs [11] (the details are in Appendix A).

Generally, we run two types of experiments: [12]
CrossValid. 5-Fold cross-validation, where we stratify on language and topics. We group the clips according to language and topic, and then we first split a 15% test set. Based on the other 85%, we apply standard k-fold CV.

Leave Topic Out. Here, we train on data of 5 topics and use the 6th topic as a test set to evaluate the transfer capabilities. During training, we employ early stopping on an evaluation set consisting of the 5 topics used for training. Thus, no information on the left-out topic is available during the training.

Early Stopping, Threshold Selection, and Evaluation. We compute the Area Under the Receiver Operating Characteristic Curve (AUC) (Marcum, 1960) score for each label after each epoch on the evaluation set and apply the Youden's J statistic (Youden, 1950) to find the optimal decision threshold for each label. We use the average F1 score over the three labels to decide on the early stopping. We then apply the selected threshold to the test set predictions. We report the macro F1

---

[11]https://huggingface.co/docs/peft/index
[12]For all our experiments, we leverage the Huggingface library https://huggingface.co

	FCW	FCN	OPN
F7B	0.889 ± 0.003	0.757 ± 0.008	0.823 ± 0.007
L3B	0.898 ± 0.002	0.772 ± 0.008	0.833 ± 0.004
M7B	0.891 ± 0.006	0.765 ± 0.005	0.829 ± 0.008
XLM	$\mathbf{0.899 \pm 0.002}$	$\mathbf{0.776 \pm 0.007}$	$\mathbf{0.836 \pm 0.004}$

Table 3: *CrossValid* F1-scores for the 4 different models and each label. The overall score is the macro F1 score. The best score in each row is in bold.

Left-Out-Topic:	F7B	L3B	M7B	XLM
European Union	0.765	0.787	0.777	0.793
League of Legends	*0.690*	*0.707*	*0.720*	*0.721*
Migration	0.726	0.768	0.766	0.763
Views about US soc.	**0.784**	**0.796**	**0.790**	**0.798**
US Elections	0.740	0.763	0.738	0.768
War in Ukraine	0.752	0.766	0.770	0.753

Table 4: *Leave Topic Out* experiments for the label detection. We report the macro F1 score for each Model and left-out topic. The best score in each column is in bold, and the worst score is in italics.

scores of the test set [13].

4.2 Results

CrossValid Results Table 3 presents the F1 scores achieved by the four models for each label in the Cross-Validation (CV) setting. Overall, we observe that the models perform consistently across the labels, with the FCW label achieving the highest F1 scores, averaging around 0.89. This is likely due to the high prevalence of this label in the training data. The OPN label also performs well, with scores ranging from 0.823 to 0.836. This indicates that the models can effectively distinguish opinions, likely due to their distinct linguistic patterns. On the other hand, the FNC label shows slightly lower performance, with F1 scores ranging from 0.757 to 0.776. This reflects the label's relatively lower prevalence and greater ambiguity in the data. Among the models, XLM-Roberta-Large (XLM) achieves the best overall performance, with the highest scores for FCW (0.899), FNC (0.776) and OPN (0.836). This demonstrates that pre-trained multilingual models fine-tuned with sufficient context can excel in multi-label classification tasks.

Leave Topic Out Results Table 4 presents the overall F1 scores for the Leave-Topic-Out experiments. As expected, performance varies depending on the left-out topic, with models performing better on topics closer to the training distribution.

The General views about US society topics achieve the highest scores across all models, ranging from 0.784 to 0.798, likely due to its broader content and overlap with other training topics. The League of Legends topic consistently shows the lowest scores, ranging from 0.690 to 0.721, reflecting its distinct domain and vocabulary compared to the other topics. Performance is moderate for the Migration and War in Ukraine topics (0.726 to 0.770), indicating that the models can generalize moderately well to topics that share argumentative or factual patterns with the training data. European Union and US Elections show slightly higher variability, with scores ranging from 0.740 to 0.793, suggesting that performance depends on the linguistic and contextual overlap with the training set.

5 Discussion & Conclusion

In this study, we introduced ViClaim, a dataset of 1,798 annotated video transcripts spanning three languages (English, German, Spanish) and six topics, including socio-political and non-political content. The dataset employs a multi-label annotation approach, distinguishing between check-worthy and non-check-worthy facts and opinions. We fine-tuned state-of-the-art multilingual models, achieving strong performance in cross-validation (macro F1 up to 0.896). However, generalization to unseen topics remains challenging, particularly for non-political domains like League of Legends, highlighting the need for more advanced approaches to domain adaptation and contextual understanding in video-based misinformation detection. By releasing ViClaim and establishing strong baselines, we aim to drive progress in multi-modal misinformation detection. This dataset addresses a critical research gap and serves as a foundation for developing tools to combat misinformation in the increasingly dominant medium of video.

Acknowledgments

Limitations

Manual Video Selection. While manual selection has its limitations, it was the most practical option given the complexity of identifying claim-rich

---

[13]We use the soft version of https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.f1_score.html using the sample weight argument to input the soft-labels

videos and the limitations of automated approaches. Although we relied on structured keyword searches to guide our selection, the process inevitably involved subjective choices by the researchers. This introduces potential biases in the dataset, such as overrepresenting certain perspectives or content types. We acknowledge that our approach was not fully systematic, but we prioritized ensuring that the selected videos aligned with the target topics and contained clear claims. By documenting our process and its limitations, we aim to provide transparency and offer a basis for future, more structured improvements in video selection methodologies.

Agreement Scores. The annotation process encountered challenges due to the subjective nature of claim detection. Disagreement among annotators was not uncommon, as reflected in moderate Krippendorff's α scores. Although MACE normalization was used to prioritize labels from more reliable annotators, this approach depends on the assumption that these annotators are consistently accurate across all contexts. Additionally, the multi-label annotation framework adds some complexity, especially for sentences with overlapping claims or mixed content, which should be considered when interpreting and applying the dataset to downstream tasks. We note that our agreement scores show a medium agreement according to the standard interpretation. However, this does not imply that our annotations are of low quality, rather it highlights the ambiguity of our task. Thus, future research can work on modeling the uncertainty inherent in such a complex task.

No multimedia analysis. Currently, the focus lies on the transcripts only. We have not yet analyzed the video and audio features of our corpus. This is part of future work and includes user comments and other meta-data in the analysis of the videos.

Topic Diversity We recognize that the dataset's topical focus does not encompass the full spectrum of claim-rich domains (e.g., health, climate, or science misinformation). Expanding the dataset to include more diverse topics is a natural next step to improve its generalizability and broaden its applicability to other real-world challenges. However, the current selection reflects our prioritization of multilingual socio-political discourse and domain transfer analysis as key research goals for this release."

Challenge of Dataset The high scores of the classifiers show that the dataset is not highly challenging. However, we are aware that in our field (machine learning and NLP), there is a tendency to require a dataset to pose a challenge so that one can create shared tasks or benchmarks around them. However, our dataset aims to be useful in downstream tasks regarding misinformation detection. Furthermore, the challenge of the datasets lies in domain transfer, where there is a large gap between the F1 score and the in-domain scenario.

Ethical Considerations

Compliance with YouTube Terms and Copyright. We collected videos exclusively from YouTube and took care to comply with YouTube's Terms of Service (ToS). Each video had a publicly visible URL at the time of selection and a median view count of 19097. We do not redistribute any copyrighted audio-visual content. Instead, we release only the video IDs and the start-to-end timestamps for each annotated sentence. We also provide code that lets researchers recreate the transcripts from these references. The original video materials remain on YouTube. Researchers wishing to access them must comply with YouTube's ToS. Our dataset neither hosts nor reproduces the videos themselves, thereby avoiding copyright violations.

Privacy of Speakers and Personal Data Mitigation. Our dataset does not include any direct personal identifiers. We do not disclose private information (e.g., personal addresses, contact details) nor do we release the audio or video footage. We also employed Named Entity Recognition (NER) checks in English, German, and Spanish to ensure the transcripts do not unintentionally reveal non-public personally identifiable information. In instances where sensitive details might have appeared, we would have masked them. Ultimately, we found that named entities refer only to the speakers themselves (who voluntarily published their videos), public figures, or fictional characters (e.g., from "League of Legends"). Therefore, no non-public personal data is disclosed.

Informed Consent. Since the nature of the task is to watch videos with partially extreme views, we informed the annotators about this and gave them the option to opt out of the task. In one case, an annotator opted out of the annotation task after the trial run. We compensated the annotator for the 1

hour they spent on the trial run and discarded their annotations.

Reputation Damage. There is a chance that there is reputational damage to the speakers in the videos being included in a dataset about misinformation. However, we only annotate whether they claim something and not whether their claims constitute misinformation.

Bias of Annotators and Researchers. The choices of the researchers and annotators influence which topics and claims are researched. Thus, there is an implicit bias towards certain topics. For this, we let the annotators fill out a questionnaire to ask about their stances on the various topics in the videos, which can be used to understand the bias.

Benefits outweigh the potential harm. We note that spreading misinformation in videos is a highly relevant and important topic. Thus, the benefits of investigating and developing analysis methods to counteract the spread of misinformation outweigh the potential harm done to the speakers.

References

AI@Meta. 2024. Llama 3 model card.

Firoj Alam, Alberto Barrón-Cedeño, Gullal S Cheema, Sherzod Hakimov, Maram Hasanain, Chengkai Li, Rubén Míguez, Hamdy Mubarak, Gautam Kishore Shahi, Wajdi Zaghouani, et al. 2023. Overview of the clef-2023 checkthat! lab task 1 on check-worthiness in multimodal and multigenre content. *Working Notes of CLEF*.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. 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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A benchmark dataset of check-worthy factual claims. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):821–829.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. In *European Conference on Information Retrieval*, pages 499–507. Springer.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Combating online misinformation videos: Characterization, detection, and future directions. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 8770–8780, New York, NY, USA. Association for Computing Machinery.

Sarmad Chandio, Muhammad Daniyal Pirwani Dar, and Rishab Nithyanand. 2024. How audit methods impact our understanding of youtube's recommendation systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 241–253.

Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. 2021. On the role of images for analyzing claims in social media. In *Proceedings of the 2nd International Workshop on Cross-lingual Event-centric Open Analytics co-located with the 30th The Web Conference (WWW 2021), Ljubljana, Slovenia, April 12, 2021 (online event due to COVID-19 outbreak)*, volume 2829 of *CEUR Workshop Proceedings*, pages 32–46. CEUR-WS.org.

Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022. MM-claims: A dataset for multimodal claim detection in social media. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 962–979, Seattle, United States. Association for Computational Linguistics.

Hyewon Choi and Youngjoong Ko. 2022. Effective fake news video detection using domain knowledge and multimodal data fusion on youtube. *Pattern Recognition Letters*, 154:44–52.

Christos Christodoulou, Nikos Salamanos, Pantelitsa Leonidou, Michail Papadakis, and Michael Sirivianos. 2023. Identifying misinformation on youtube through transcript contextual analysis with transformer models. *arXiv preprint arXiv:2307.12155*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2021. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Mudit Dhawan, Shakshi Sharma, Aditya Kadam, Rajesh Sharma, and Ponnurangam Kumaraguru. 2022. Game-on: Graph attention network based multimodal fusion for fake news detection. *arXiv preprint arXiv:2202.12478*.

Mark Dingemanse and Andreas Liesenfeld. 2022. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5614–5633, Dublin, Ireland. Association for Computational Linguistics.

Subhabrata Dutta, Rudra Dhar, Prantik Guha, Arpan Murmu, and Dipankar Das. 2023. A multilingual dataset for identification of factual claims in indian twitter. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '22, page 88–92, New York, NY, USA. Association for Computing Machinery.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

A. Goldberg and F. Marquart. 2021. "that's just, like, your opinion" – european views on argumentation and debate. *European Journal of Argumentation*, 35(2):45–67.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Rui Hou, Verónica Pérez-Rosas, Stacy Loeb, and Rada Mihalcea. 2019. Towards automatic detection of misinformation in online medical videos. In *2019 International conference on multimodal interaction*, pages 235–243.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring misinformation in video search platforms: An audit study on youtube. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.

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, Online. Association for Computational Linguistics.

J. Marcum. 1960. A statistical theory of target detection by pulsed radar. *IRE Transactions on Information Theory*, 6(2):59–267.

Nicholas Micallef, Marcelo Sandoval-Castañeda, Adi Cohen, Mustaque Ahamad, Srijan Kumar, and Nasir Memon. 2022. Cross-platform multimodal misinformation: Taxonomy, characteristics and detection for textual posts and videos. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):651–662.

Priyank Palod, Ayush Patwari, Sudhanshu Bahety, Saurabh Bagchi, and Pawan Goyal. 2019. Misleading metadata detection on youtube. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41*, pages 140–147. Springer.

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.

Kostantinos Papadamou, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. 2022. "it is just a flu": Assessing the effect of watch history on youtube's pseudoscientific video recommendations. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):723–734.

Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, Yiannis Kompatsiaris, and Denis Teyssou. 2018. Invid fake video corpus v2.0.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Noushin Salek Faramarzi, Fateme Hashemi Chaleshtori, Hossein Shirazi, Indrakshi Ray, and Ritwik Banerjee. 2023. Claim extraction and dynamic stance detection in covid-19 tweets. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 1059–1068, New York, NY, USA. Association for Computing Machinery.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, Alex Nikolov, Wajdi Zaghouani, Preslav Nakov, and Anna Feldman. 2021. Findings of the nlp4if-2021 shared tasks on fighting the covid-19 infodemic and censorship detection. *arXiv preprint arXiv:2109.12986*.

A Srinivasacharlu. 2020. Using youtube in colleges of education. *Shanlax International Journal of Education*, 8(2):21–24.

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Empowering the fact-checkers! automatic identification of claim spans on Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7701–7715, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pius von Däniken, Jan Milan Deriu, and Mark Cieliebak. 2023. Zhaw-cai at checkthat! 2023: Ensembling using kernel averaging. In *14th Conference and Labs of the Evaluation Forum (CLEF), Thessaloniki, Greece, 18-21 September 2023*, pages 534–545. CEUR Workshop Proceedings.

Hazaymeh Wafa'A and Mohamad Ahmad Saleem Khasawneh. 2024. The correlation between the use of youtube short videos to enhance foreign language reading and writing proficiency, and the academic performance of undergraduates. *Eurasian Journal of Educational Research*, 109(109):1–13.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

W. J. Youden. 1950. Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Xichen Zhang and Ali A Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025.

A Hyperparameters

The hyperparameters were selected based on manual trial and error. The budget for GPU computation was limited, and a grid search approach would have been prohibitively expensive. For all models, we used an 8-bit ADAM optimizer via block-wise quantization (Dettmers et al., 2021). The experiments were run on a GPU cluster with 8 NVIDIA H200 GPUs, and the total experimentation time was approximately 100 GPU hours.

XLM-Robertal-Large (XLM). Conneau et al. (2020) pre-trained an 550M parameter encoder-transformer on 100 languages. For claim detection, we used a learning rate of $5e - 04$, a batch size of 32 with a gradient accumulation of 8.

Falcon-7B (F7B). Almazrouei et al. (2023) pre-trained a 7B decoder-only model on 1.5T tokens of the RefinedWeb corpus (Penedo et al., 2023). The main languages that Falcon performs well on are English, German, Spanish, and French, which cover our use case well. We applied QLorA (Dettmers et al., 2024) with rank 16, alpha 32, a dropout of 0.05, and a learning rate of $5e - 04$. For claim detection we used a batch size of 64 with a gradient accumulation of 8.

Mistral-7B (M7B). Jiang et al. (2023) pre-trained a 7B parameter decoder-only model. The details of the training data used are undisclosed. We applied QLorA (Dettmers et al., 2024) with rank 16, alpha 8 a dropout of 0.05 and a learning rate of $5e - 04$. For claim detection we used a batch size of 64 with a gradient accumulation of 8.

LLama3.2-3B (L3B). AI@Meta (2024) pre-trained an 8B parameter decoder-only model trained on an undisclosed set of 15T tokens of web

data. We applied QLorA (Dettmers et al., 2024) with rank 64, alpha 16, and dropout of 0.1, and a learning rate of $5e - 04$. For claim detection we used a batch size of 64 with a gradient accumulation of 8.

B Dataset Distribution Overview

Table 5 presents the distribution of videos and annotated sentences across topics and languages. Initially, we distributed an equal number of 600 videos to each annotation group, as outlined in subsection 3.5.

Given the substantial volume of 1,800 videos, it was not feasible to manually verify the transcripts generated for each video. Instead, annotators were provided with a feature in the annotation tool to flag issues if they encountered problems with the videos. Some transcripts were improperly generated during the transcription pipeline, leading to significant errors, such as missing most of the spoken text or being transcribed into the wrong language. These issues contributed to the irregular distribution displayed in Table 5.

To address these challenges, we sought to collect new video data with similar content to replace the problematic entries. However, this effort resulted in a slightly reduced and inconsistently distributed collection of annotated videos compared to the initial dataset plan. Table 5 reflects these adjustments and the final dataset composition.

C Ambiguous Sentences

Table 6 shows a set of ambiguous sentences where multiple labels are applicable. These are examples where the annotators disagreed. In most cases, disagreement in the labels stemmed from such sentences, where one annotator only selected one class. In contrast, another annotator selected the other class, but both classes would have been applicable. Thus, for this reason, we opted to work with soft-labels to cover this kind of ambiguity.

D Exit Form Output

After completing the annotation process, we distributed an exit form to our 12 annotators to gain a pseudonymized view of their biases and perspectives across the six annotation topics. Each annotator used their unique ID to provide responses anonymously. The exit form included questions about demographics, social status, political views, and topic-specific questions to capture opinions and insights. Annotators were allowed to take a side, remain neutral, or leave certain questions unanswered if they preferred not to disclose their opinions. Most of the annotators were students between the age of 25-34 years and native German speakers. Regarding topics for US Elections, we see that most annotators would rather support Biden but still think there should be a more suitable candidate. For the topic of War in Ukraine, we see that most annotators would side with Ukraine and see Russia as the aggressor. For the Migration topic, annotators' opinions were evenly distributed on whether migrants are responsible for increased criminality. Many were unsure about whether migrants are beneficial for their host countries, with some agreeing that migrants fill economic gaps. For the topic of General views about the US society, it is also the annotator's view that the educational system in the US is not quite as good for the broad society. However, they agree that the US is most likely still the most dominant military power in the world. Annotators seem not to be biased against or for US citizens in general. For the topic of the European Union, most annotators agree that the EU is an organization supporting a better world. For the topic of League of Legends, annotators can not really relate due to not having ever played that game, but they would agree that it is rather addictive. We conclude from the exit form, that the annotations made in the dataset would have a minor bias, as described in the summarized results of the form to the corresponding topics. The great majority of annotators stated, that they would do the annotation task again and even would recommend a friend to participate.

E Closed Source Models

In a preliminary step, we evaluated three proprietary systems Grok2 [14], o4-mini [15] and o3-mini [16] on our classification task. For this, we prompted each model with the task descriptions from our guidelines and added the examples from the paper to the prompt for in-context learning. Their F1 scores are far below the fine-tuned models listed in the table 7. The best performing models for each respecting label only achieved an F1 score for FCW: 0.7797, F1 score for FNC: 0.5559, and

---

[14] https://x.ai/news/grok-2
[15] https://platform.openai.com/docs/models/o4-mini
[16] https://platform.openai.com/docs/models/o3-mini

	en		de		es	
Topic	**Videos**	**Sentences**	**Videos**	**Sentences**	**Videos**	**Sentences**
US Elections 2024	97	1111	107	1007	103	692
Societal issues in the US	100	1401	98	1133	89	575
War in Ukraine	95	922	109	1009	121	915
Migration	96	1211	112	1235	100	714
League of Legends	97	1075	90	954	90	760
European Union	100	879	98	790	96	733
Total	585	6599	614	6128	599	4389
Total Videos			1798			
Total Sentences			17116			

Table 5: Overview of the dataset distribution of videos in each language and for each topic.

Sentence	Labels	Explanation
Russia and China are states that are confident enough to withstand all kinds of pressure.	FNC or OPN	Even if it looks like a factual claim, this is an opinion per guidelines, as it is speculative, subjective, and lacks verifiable information.
Brexit has failed.	FNC or OPN	Even if it looks like a factual claim, it is hard to proof and seems very subjective. Britain is still a former member of the EU, not a current one.
Trump was glitching so badly looking bloated and haggard, sweaty and disoriented, gripping the lectern like his life depended on it.	FCW, FNC or/ and OPN	This is a fact-checkworthy claim, as we can verify how Trump appeared in the video. However, the additional description is highly subjective, qualifying it as an opinion as well. Additionally it is questionable if this is really from public interest how Trump looked on a certain video and could therefore also pass as factual not checkworthy claim.
Maybe because from the time that we get into school, women are being told not to have children, not to aspire to family.	FCW, FNC or OPN	This statement could fit any of the three labels. While it is conveyed as a factual statement and could potentially be verified against school guidelines (though unlikely), it leans toward being a fact-non-checkworthy claim. However, given its opinionated and subjective nature, it could also be classified as an opinion.
Well, let's just say that it was a different time when Vandal Jacks was made.	FNC or OPN	This statement is presented as a fact but lacks checkworthiness due to insufficient information. Its speculative nature also makes it suitable to classify as an opinion.
You underestimated the capability and the fighting spirit of the brave ukrainian people.	FCW, FNC or/ and OPN	This statement is presented as a fact but is speculative, with the second part clearly an opinion. It could be classified as fact-checkworthy, fact-non-checkworthy, or opinion, depending on the source and context. Interviews could also be checked to verify whether Russian generals were initially overly confident.
You misjudged the international community, who has universally condemned your actions.	FCW, FNC or OPN	This statement could be classified as fact-checkworthy, fact-non-checkworthy, or opinion, given its speculative nature. The classification depends significantly on the context and the identity of the person making the claim.

Table 6: Examples of sentences where multiple interpretations of tags are valid due to ambiguity.

	F1		
Model	**FCW**	**FNC**	**OPN**
o3-mini	**0.7797**	0.5408	**0.7789**
4o-mini	0.6498	**0.5559**	0.7509
Grok 2 (closed source)	0.6650	0.5490	0.6300

Table 7: Scores for closed source models on the classification task.

F1 score for OPN: 0.7789. Thus, we abandoned closed-source models for this task and focused on fine-tuning classification models.

F Comparison of Written vs. Spoken Language

To showcase that spoken and written language differ, we used the XLM-Roberta-based classifiers provided by (von Däniken et al., 2023), which are fine-tuned on the CheckThat 2023 Twitter data for claim check-worthiness classification (Alam et al., 2023). For the in-domain data (i.e., test set consisting of Tweets), it achieved an F1 score of 0.693 (with equilibrated precision and recall); however, on our data, it only achieved an F1 score of 0.32, with precision at 0.72, and recall at 0.2. We also fine-tuned a ModernBERT (Warner et al., 2024),

which achieved an F1 score of 0.46, with a precision of 0.54 and a recall of 0.40. This shows that the existing written text datasets are inadequate for our usecase.
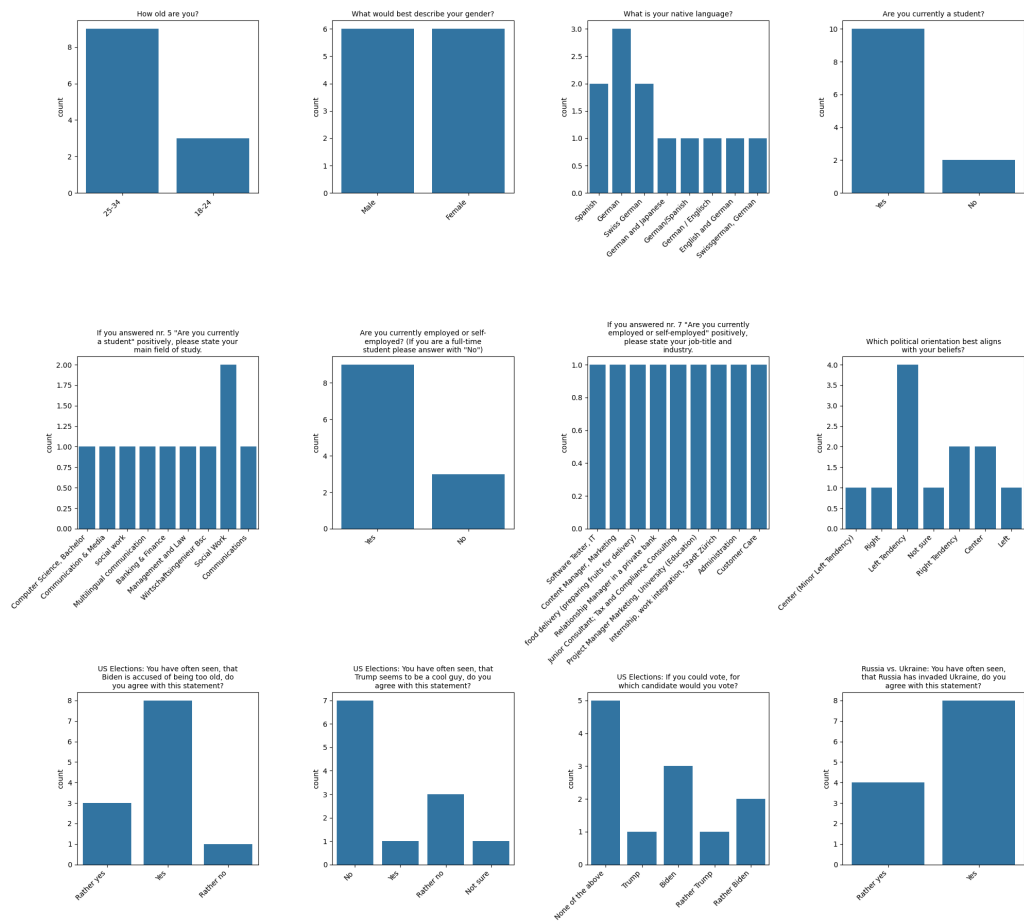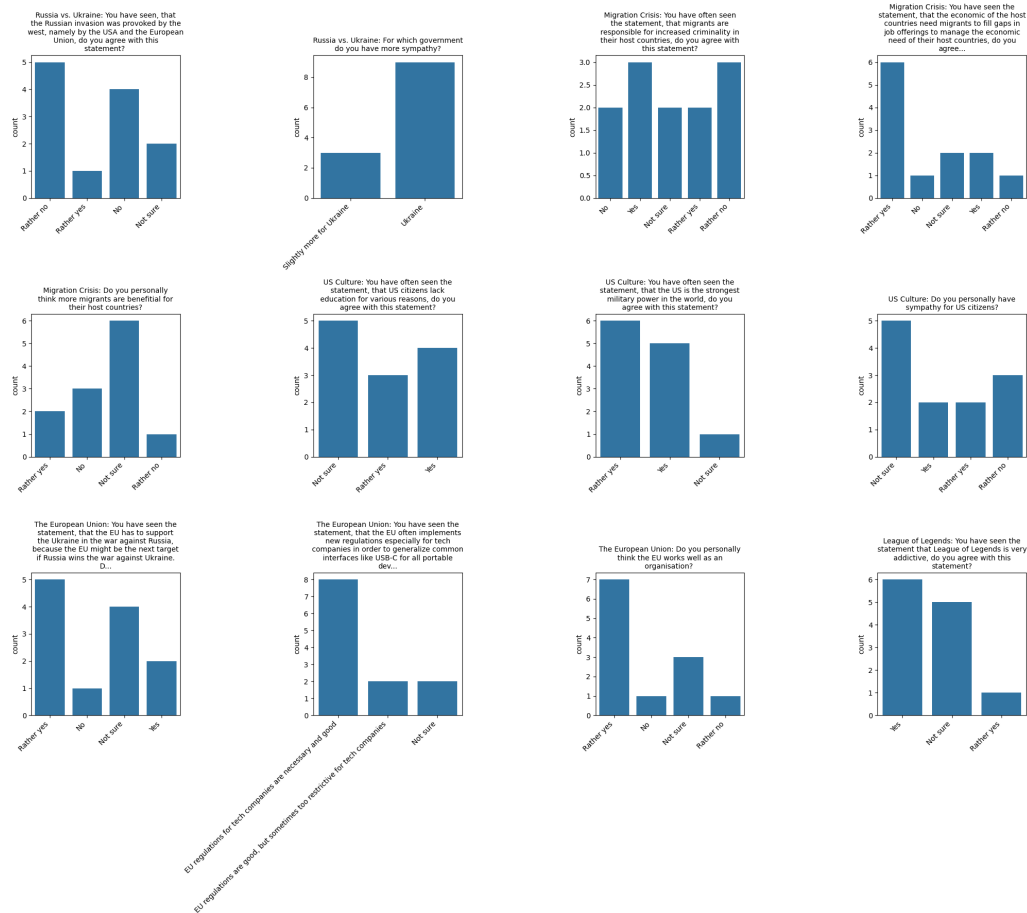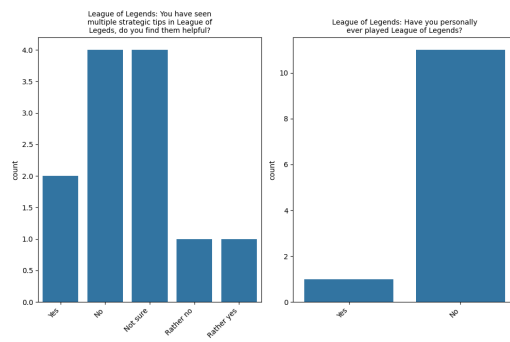
Figure 3: Form questions 1 to 12

Figure 4: Form questions: 13 to 24

Figure 5: Form questions 25 to 29