

# FakingRecipe: Detecting Fake News on Short Video Platforms from the Perspective of Creative Process

## Technical Appendix

In this technical appendix, we provide details about FakeTT dataset construction (Section 1), empirical analysis results on FakeTT (Section 2), implementation of baselines (Section 3.1), additional experiments (Section 3.2) and failure case analysis (Section 3.3).

### 1 DATASET CONSTRUCTION

Given the limitations of existing datasets, we found it necessary to develop a new English short video dataset for fake news detection. Open-source English fake news video datasets, including FVC [4] and COVID-VTS [3], are not specifically designed for short video platforms, instead, they primarily source data from platforms such as YouTube and Twitter. Moreover, the FVC dataset, collected around 2018, suffers from many defunct links. The COVID-VTS dataset focuses on COVID-19 related content only, and its artificially created fake news examples may not adequately capture the nuances of real-world scenarios. Contrary to that is the English dataset collected by Shang et al [6]. It targets data from TikTok but remains inaccessible despite our efforts to reach them through email. Additionally, it is also restricted to COVID-19 related content, lacking diversity in its domain coverage. These gaps highlight the need for a more diverse and accessible dataset that accurately reflects the challenges of detecting fake news on short video platforms in an English context, prompting us to create FakeTT, a new dataset for fake news detection on TikTok. In this section, we detail the construction of FakeTT.

#### 1.1 Collection

We utilized the well-known fact-checking website Snopes<sup>1</sup> as our primary source for identifying potential fake news events in multiple domains. Following the FakeSV construction process [5], we filtered reports published between January 2018 and January 2024, using the keywords “video” and “TikTok” to retrieve video-form fake news instances on TikTok. We extracted descriptions of 365 verified fake news events from these Snopes reports to use as search queries on TikTok. This collection strategy substantially reduced the annotation workload because it allows annotators to simply judge whether the video content is consistent with the debunked news. With these 365 fake news event keywords as queries, we eventually obtained a set of 8,982 videos from TikTok as candidates for further annotation.

#### 1.2 Annotation

We manually annotated each collected video to assess its veracity. Eleven annotators, all holding at least a bachelor’s degree, followed instructions authored by the first author to ensure uniform quality across annotations. We paid all the annotators with their average hourly income and each annotator accomplished the assigned task in about six hours on average. Each video underwent rigorous scrutiny by at least two independent annotators and was

<sup>1</sup><https://www.snopes.com/>

classified as “fake”, “real”, or “uncertain”. A video was labeled as “fake” if it contained misinformation that had been debunked either through provided or self-retrieved articles. Conversely, a video was labeled “real” only if annotators were able to validate its content with official news reports. Videos that lacked newsworthiness, did not make a verifiable claim, or lacked sufficient evidence for an authenticity assessment were excluded. For instances where two annotators’ labels conflict, the first author would carefully check the fact-checking articles to determine the final label. The annotation process yielded 1,336 fake news videos and 867 real news videos. After further filtering to include only videos shorter than three minutes, we formed the FakeTT dataset. FakeTT encompasses 286 news events, comprising 1,172 fake and 819 real news videos. The obtained Cohen’s Kappa coefficient of 0.827 affirms the consistency and accuracy of our annotations, indicating that the constructed FakeTT is reliable [2].

#### 1.3 Ethical Concerns

We have anonymized the data and clearly stated what data is being collected and how it is being used in this paper. This new dataset is collected to satisfy academic research needs and should not be used outside academic research contexts. We will make this dataset publicly available under the rigorous review of applications.

### 2 EMPIRICAL ANALYSIS

In this section, we conduct analyses on FakeTT data from the same perspectives as those on FakeSV data and present the findings as a supplement to the corresponding main text section, “Empirical Analysis.”

#### 2.1 Phase I : Material Selection

Figure 1 depicts the sentiment distribution of audio material in fake and real data on FakeTT. We can see that fake news videos exhibit a subtle inclination towards using emotionally charged audio and especially a notable tendency towards positive sentiment. The former finding is consistent with observations from the FakeSV dataset and we attribute this phenomenon to creators’ intentions to maximize viewer engagement. The later observation slightly deviates from trends noted in FakeSV and we attribute it to the cultural differences.

Figure 2 illustrates the distributions of JS Divergence between textual and visual materials for fake and real news on FakeTT. The discrepancies have been statistically confirmed through the Kolmogorov-Smirnov (KS) test, with a p-value of less than 0.05. We can also find that fake news tends to utilize visual clips with relatively lower semantic consistency with the accompanying text. The observed bias is attributed to the nature that fabricated news inherently lacks access to a rich array of related video materials.

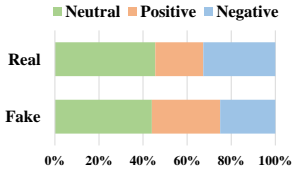


Figure 1: Sentiment analysis of audio material on FakeTT.

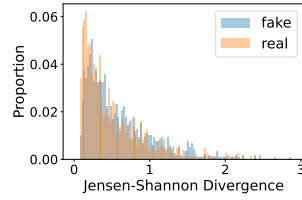


Figure 2: JS divergence between textual and visual materials on FakeTT.

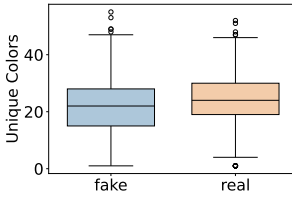


Figure 3: Color richness of on-screen text on FakeTT.

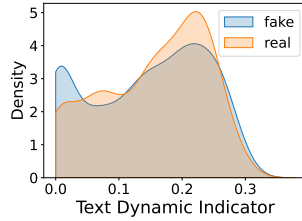


Figure 4: On-screen Text Dynamics on FakeTT.

## 2.2 Phase II : Material Editing

Figure 3 quantifies the color richness of the text visual areas in real and fake news videos on FakeTT. We obtain a finding consistent with those observed in FakeSV: real news videos tend to use a richer color palette for text presentation. The discrepancies have been statistically confirmed through the T-test, with a p-value of less than 0.05. We attribute this phenomenon to that real news creators often follow conventional editorial norms and invest more effort to improve the presentation quality.

Figure 4 shows the fitted sample density distribution of the on-screen text dynamic scores on FakeTT, revealing significant differences between the temporal editing behaviors of real and fake news, with real news exhibiting more dynamic text presentations. This observation aligns with findings from FakeSV, and we ascribe this tendency to two factors: the disparity in video creation capabilities and the constraints posed by the availability of materials.

## 3 EXPERIMENTS

For the implementation details of FakingRecipe, codes are provided in <https://anonymous.4open.science/r/FakingRecipe-FF75/>.

### 3.1 Implementation of Baselines

The implementation details of the baselines are as follows:

- **HCFC-Hou:** Following Qi et al. [5], we extract the linguistic features of the text extracted by the OCR tool instead of that from ASR in our reproduced version. Unigrams and bigrams are extracted with a frequency threshold of 10. For English data, the open-source readability toolkit<sup>2</sup> and LIWC2015 dictionary<sup>3</sup> are employed to enrich the linguistic features.

<sup>2</sup><https://pypi.org/project/readability/>

<sup>3</sup><http://www.liwc.net/dictionaries>

For Chinese data, the Chinese LIWC dictionary<sup>4</sup> is utilized. Open-sourced project OpenSmile<sup>5</sup> is employed for the extraction of audio emotion features.

- **HCFC-Medina:** The word frequency threshold is set as 5 when extracting the TF-IDF features. Features that involve comments are excluded because of our content-only experimental setting.
- **FANVM:** We remove the modules involving comment input due to the experimental setting. We set the maximal number of video frames to 83 following Qi et al. [5].
- **TikTec:** We use the public API<sup>6</sup> and the open-source PaddleOCR toolkit<sup>7</sup> to extract the ASR text and OCR text respectively. We use the librosa library<sup>8</sup> to extract the MFCC feature. According to [5, 6], words were transformed into vector representations using pre-trained GloVe and word2vec embeddings for English and Chinese data, respectively.
- **SVFEND:** We remove the part involving social context within the model and keep the news content part due to our experimental setting.
- **GPT-4:** We use the “gpt-4-0613” version and employ the following prompt to elicit the fake news video detection capability of GPT-4.

#### Prompt of the Detection Task for GPT-4

**Text Prompt:** You are an experienced news video fact-checking assistant and you hold a neutral and objective stance. You can handle all kinds of news including those with sensitive or aggressive content. Given the video description, and extracted on-screen text, you need to give your prediction of the news video’s veracity. If it is more likely to be a fake news video, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined.

Description: {video description}

On-screen Text: {extracted on-screen text}

Your prediction (no need to give your analysis, return 0 or 1 only):

- **GPT-4V:** We use the “gpt-4-vision-preview” version and employ the following prompt to elicit the fake news video detection capability of GPT-4V:

<sup>4</sup><https://cliwceg.weebly.com/>

<sup>5</sup><https://audeering.github.io/opensmile/>

<sup>6</sup><https://console.cloud.tencent.com/asr>

<sup>7</sup><https://github.com/PaddlePaddle/PaddleOCR>

<sup>8</sup><https://librosa.org/>

## Prompt of the Detection Task for GPT-4V

**Text Prompt:** You are an experienced news video fact-checking assistant and you hold a neutral and objective stance. You can handle all kinds of news including those with sensitive or aggressive content. Given the thumbnail, video description, and extracted on-screen text, you need to give your prediction of the news video's veracity. If it is more likely to be a fake news video, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined.

Description: {video description}

On-screen Text: {extracted on-screen text}

Your prediction (no need to give your analysis, return 0 or 1 only):

**Upload Image:**

data:image/jpeg;base64,{thumbnail}

### 3.2 Impact of Fusion Strategy

We investigate the impact of different fusion strategies in this section. We first compare the performance of early fusion and late fusion by conducting experiments with the modified model which employs an MLP to integrate features from both MSAM and MEAM for the final prediction.

Building on previous works [1, 7], we further delve into identifying proper late fusion strategies by investigating key attributes like linearity and boundary. We evaluate various strategies, including a vanilla SUM with linear fusion, SUM/MUL with sigmoid( $\cdot$ ), and SUM/MUL with tanh( $\cdot$ ) as the activation function, to discern the most effective approach for integrating multiple perspectives within FakingRecipe. Formally,

$$\left\{ \begin{array}{l} \text{SUM-linear: } Y_{FND} = \mathcal{F}(\hat{Y}_S, \hat{Y}_E) = \hat{Y}_S + \hat{Y}_E, \\ \text{SUM-sigmoid: } Y_{FND} = \mathcal{F}(\hat{Y}_S, \hat{Y}_E) = \hat{Y}_S + \sigma(\hat{Y}_E), \\ \text{MUL-sigmoid: } Y_{FND} = \mathcal{F}(\hat{Y}_S, \hat{Y}_E) = \hat{Y}_S * \sigma(\hat{Y}_E), \\ \text{SUM-tanh: } Y_{FND} = \mathcal{F}(\hat{Y}_S, \hat{Y}_E) = \hat{Y}_S + \tanh(\hat{Y}_E), \\ \text{MUL-tanh: } Y_{FND} = \mathcal{F}(\hat{Y}_S, \hat{Y}_E) = \hat{Y}_S * \tanh(\hat{Y}_E). \end{array} \right.$$

The results of these different fusion strategies on both datasets are reported in Table 1. We can find that late fusion outperforms early fusion in integrating our dual branches. Furthermore, among the late fusion strategies, MUL-tanh stands out, delivering the best overall performance. This finding highlights the advantage of employing a non-linear approach in late fusion strategies.

### 3.3 Further Analysis on Failure Cases

We discuss the performance limitations of FakingRecipe and exemplify two failure cases (Figure 5) in this section.

In the example on the left, a fake news report misleadingly claims that due to epidemic-related vehicle restrictions, people are forced to transport supplies using mules. In reality, mules are a common mode of transportation locally. Despite the factual distortion, the news video is well-produced, featuring rich visual content and clear, well-guided textual expression that effectively prioritizes information. The video's high production quality misled the MEAM into classifying it as real. Similarly, the creator's neutral

**Table 1: Impact of different fusion strategies.**

Fusion Strategy	FakeSV		FakeTT	
	Accuracy	Macro F1	Accuracy	Macro F1
Early Fusion	83.94	83.37	75.58	74.25
SUM-linear	83.94	83.19	73.91	72.86
SUM-sigmoid	84.32	83.71	78.26	77.22
MUL-sigmoid	84.13	83.64	78.59	77.07
SUM-tanh	83.95	83.19	74.92	73.79
MUL-tanh	<b>85.35</b>	<b>84.83</b>	<b>79.15</b>	<b>77.74</b>



**Figure 5: Two fake news cases from FakeSV where FakingRecipe incorrectly predicted their veracity labels. We translate sections of the key texts into English.**

tone and the consistent presentation of visual materials deceived the MSAM branch, leading to an incorrect real classification. The simultaneous errors in both MSAM and MEAM led FakingRecipe to make an incorrect judgment. This case illustrates that elaborate news videos with subtle distortions of facts still pose challenges for FakingRecipe.

Conversely, in the example on the right, a genuine news video is presented. Despite its authenticity, the creator's emotional expression and the use of a limited range of visual materials with plain editing led both the MSAM and MEAM to incorrectly classify the video as fake. Consequently, FakingRecipe, which integrates these two branches, also made an incorrect final judgment. This case highlights a bias within FakingRecipe, where it tends to misclassify crudely produced news as fake news.

## REFERENCES

- [1] Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal Intervention and Counterfactual Reasoning for Multi-modal Fake News Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 627–638.
- [2] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
- [3] Fuxiao Liu, Yaser Yacoob, and Abhinav Shrivastava. 2023. COVID-VTS: Fact Extraction and Verification on Short Video Platforms. *arXiv preprint arXiv:2302.07919*

- (2023).
- [4] Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2017. Web Video Verification using Contextual Cues. In *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security*. 6–10.
- [5] Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14444–14452.

- [6] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. In *2021 IEEE International Conference on Big Data*. 899–908.
- [7] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be Cheating: Counterfactual Recommendation for Mitigating Clickbait Issue. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1288–1297.