FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms

Peng Qi[1,2], Yuyan Bu[1,2], Juan Cao[1,2]*, Wei Ji[3], Ruihao Shui[3], Junbin Xiao[3], Danding Wang[1], Tat-Seng Chua[3]

[1] Key Lab of Intelligent Information Processing, Institute of Computing Technology, CAS
[2] University of Chinese Academy of Sciences
[3] National University of Singapore
{qipeng,caojuan,wangdanding}@ict.ac.cn, buyuyan22@mails.ucas.ac.cn,
{jiwei,dcscts}@nus.edu.sg, ruihaoshui@u.nus.edu, junbin@comp.nus.edu.sg

Abstract

Short video platforms have become an important channel for news sharing, but also a new breeding ground for fake news. To mitigate this problem, research of fake news video detection has recently received a lot of attention. Existing works face two roadblocks: the scarcity of comprehensive and large-scale datasets and insufficient utilization of multimodal information. Therefore, in this paper, we construct the largest Chinese short video dataset about fake news named FakeSV, which includes news content, user comments, and publisher profiles simultaneously. To understand the characteristics of fake news videos, we conduct exploratory analysis of FakeSV from different perspectives. Moreover, we provide a new multimodal detection model named SV-FEND, which exploits the cross-modal correlations to select the most informative features and utilizes the social context information for detection. Extensive experiments evaluate the superiority of the proposed method and provide detailed comparisons of different methods and modalities for future works. Our dataset and codes are available in https://github.com/ICTMCG/FakeSV.

(a) (b)

Figure 1: Typical cases of fake news videos. Figure (a) shows a fake news video of which the content of some frames have been tampered. The spliced region is marked in red line. (b) shows a fake news video that is contextually inappropriate, where the comments provide key clues.

Introduction

With the prevalence of short video platforms (*e.g.,* Tiktok), they have become an important channel for news sharing (Walker and Matsa 2021). Besides professional news outlets, ordinary users also upload news videos happening around them. However, this openness have led to such platforms becoming a new breeding ground for fake news. Video modality is more powerful in spreading fake news than other modalities, and thus will exacerbate the influences of fake news (Sundar, Molina, and Cho 2021). Therefore, studying video-form fake news is important for detecting and intervening the spread of fake news in a timely manner.

Fake news videos refer to those that the video content and title jointly describe a piece of news that is verifiably false. Compared with traditional text-based or text-image fake news detection, fake news video detection presents unique challenges: First, video-form fake news has more modalities and thus more information. We need to select the most informative clues from multiple modalities and fuse heterogeneous information to understand the news content and detect fake news. Second, different from traditional video platforms such as YouTube, short video platforms provide advanced video-editing functions (*e.g.,* text box and virtual background) for users to conveniently modify the videos. It weakens the discriminability of the visual content in measuring its truthfulness as both fake and real news videos could be modified and exhibit editing traces. Moreover, fake news videos typically contain real news videos where only some frames or the accompanying titles have been maliciously modified to alter their meanings (as shown in Figure 1). All these characteristics make it inadequate to study fake news videos from news content only. Thus, we need to explore auxiliary social context information, such as user comments (Shu et al. 2019a) (as shown in Figure 1(b)) and user profiles (Shu, Wang, and Liu 2018), to help the detection.

Most works studying fake news focus on text-based or text-image fake news, while few works pay attention to video-form fake news. There are two major limitations of existing works: 1) **Dataset**: Although there are several datasets for fake news video detection, the majority of them are constructed on traditional video platforms and contain only a few hundred instances, and none of them provides information on news content, user comments, and user profiles simultaneously. Further, the lack of comprehensive datasets

*Corresponding author.

with rich modalities hinders the fair comparisons of existing works that use different modalities for detection. 2) **Detection methods**: Most works targeting video-based fake news detection are still in the preliminary stage of hand-crafted features. Additionally, none of them considers all modalities and thus inevitably misses some important clues.

To bridge the gap of fake news video datasets, we propose a large-scale Chinese **Fake** News **S**hort **V**ideo dataset (**FakeSV** for short) that includes both complete news content and rich social context. The abundant features in this dataset not only provide an opportunity to evaluate different approaches for fake news detection, but also help understand the diffusion of fake news and its intervention. To study the characteristics of fake news videos, we comprehensively conduct exploratory analysis of FakeSV from different perspectives including the multimodal news content, social context, and propagation, which could shed light on detection strategies.

To tackle the challenges of fake news video detection, we propose a multimodal model named **SV-FEND** (**S**hort **V**ideo **F**ak**E** **N**ews **D**etection) as a new baseline on FakeSV. The proposed model utilizes the co-attention mechanism to enhance the multimodal content representations by spotting the important features, and fuses them with the social context features by the self-attention mechanism. We conduct extensive experiments comparing the proposed model and existing baseline methods on FakeSV. The results validate the superiority of SV-FEND and provide detailed insights of different methods and modalities for future works.

Our contributions are summarized in three aspects:

- We construct the largest Chinese fake news short video dataset, namely FakeSV. This dataset contains complete news contents and rich social context, and thus can support a wide range of research tasks related to fake news. We also provide in-depth statistical analysis.
- We provide a new multimodal baseline method SV-FEND that captures the multimodal correlations to enhance the news content representations and utilizes the signals of social context to help the detection.
- We conduct extensive experiments with the proposed model and existing SOTA methods on FakeSV, which validate the superiority of the proposed method and provide detailed comparisons of different methods and modalities for future works.

Related Work

Datasets. As video-form fake news has drawn increasing attention, there have been efforts in constructing datasets to detect fake news videos. Table 1 summarizes existing datasets in aspects of richness of features, size, domain, language and sources. Because annotators need to spend more time to watch and understand the video content, video-form fake news datasets are typically smaller than text-based and text-image datasets (Murayama 2021), *e.g.,* 5 of 7 datasets only contain less than 1000 instances. Moreover, most of the small-scale datasets provide limited features and only focus on one domain. Thus they are not suitable to study fake news especially in developing detection models that generalize well to new domains. In particular, Papadopoulou et al. (2018) build the largest video-form fake news dataset. They first annotate an initial set of 380 videos and then collect related videos by searching the multi-lingual titles of these seed videos on YouTube, Tweet, and Facebook. This dataset provides information about the news content and user comments. However, these videos are collected from traditional video platform and social medias where editions of videos are not as prevalent and deep as those on emerging short video platforms, and thus could not reflect the latest challenge in fake news video detection. In addition, the publisher profiles are ignored in this dataset.

Techniques. Far behind text-image fake news detection where many well-designed neural networks have been proposed to model different types of text-image correlations (Qi et al. 2021), most works oriented for video-based fake news detection are still in the preliminary stage of hand-crafted features. As the first work of detecting fake news videos, (Papadopoulou et al. 2018) build an SVM classifier over features based on the video metadata, title linguistics, and comment credibility. (Medina et al. 2020) use the tf-idf vectors of title and comments, and comments conspiracy to make a classification. (Hou et al. 2019) firstly import emotion acoustic features and (Li et al. 2022) convert audio into text to extract linguistic features. Inspired by the rapid development of deep neural networks, (Palod et al. 2019) use LSTM to model the comment features and concatenate them with hand-crafted features. (Choi and Ko 2021) propose a well-designed model that uses pre-trained models to extract features of video frames, title, and comments. The difference in topic distribution between title and comments is used to fuse these two modalities and a topic-adversarial classification is used to guide the model to learn topic-agnostic features for good generalization. (Shang et al. 2021) use the extracted speech text to guide the feature learning of visual frames, use MFCC (Palo, Chandra, and Mohanty 2018) features to enhance the speech text, and then use a co-attention module to fuse the visual and speech information.

Although existing works have made some achievements in fake news video detection, they only utilize partial modalities that is available in the experimental dataset for detection, which inevitably misses some important clues. Moreover, as these methods are evaluated on different datasets, there is a lack of an effective comparison between them. Therefore, in this work, we propose a new multimodal detection model which considers all the involved modalities and reimplement these representative baseline methods on the proposed FakeSV dataset for a fair comparison.

Dataset Construction

Next, we describe the construction process of the FakeSV dataset as shown in Figure 2.

Data Collection

To get as much fake news as possible and collect reliable ground truth labels for fake news, we utilized fact-checking

Dataset	Features					Instances (fake/real)	Domain	Language	Source
	Video	Title	Metadata	Comment	User				
(Papadopoulou et al. 2018)	✓	✓	✓	✓		2,916/2,090	-	En,Fr,Ru, Ge,Ar	YT,TW, FB
(Palod et al. 2019)	✓	✓	✓	✓		123/423	-	En	YT
(Hou et al. 2019)	✓		✓			118/132	prostate cancer	En	YT
(Medina et al. 2020)	✓	✓		✓		113/67	COVID-19	En	YT
(Choi and Ko 2021)	✓	✓		✓		902/903	-	En	YT
(Shang et al. 2021)	✓	✓				226/665	COVID-19	En	TT
(Li et al. 2022)	✓		✓			210/490	health	Ch	BB
FakeSV (ours)	✓	✓	✓	✓	✓	**1,827/1,827**	**-**	Ch	**DY, KS**

Table 1: Summary of datasets of fake news video detection. Metadata refers to basic statistics such as # of likes/stars/comments. "-" represents open-domain. Names of sources are abbreviated for simplicity (YT: YouTube, TW: Twitter, FB: Facebook, TT: TikTok, BB: Bilibili, DY: Douyin, KS: Kuaishou).

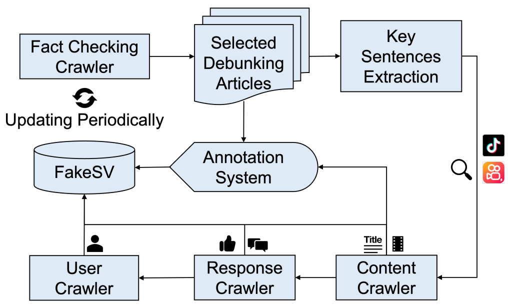
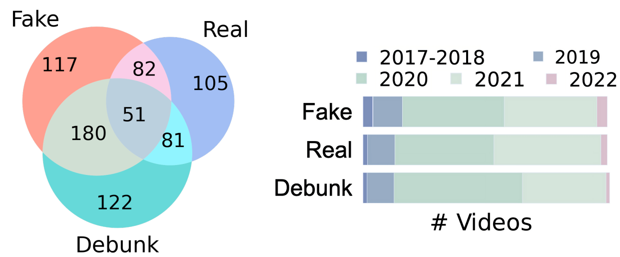


Figure 2: Flowchart of data construction process.



(a) Event (b) Year

Figure 3: Distribution of FakeSV on event and year.

Category	Fileds
Content	video, cover image, title, published time
Response	# of likes/stars/comments, top 100 comments (with reviewed time, # of likes and # of sub-comments)
Publisher	info_verified, info_introduction, current IP location, # of fans/subscribes/likes/videos and top 100 published videos' covers

Table 2: Crawled fields in FakeSV.

websites to obtain the target events. We crawled debunking articles from several official fact-checking sites between January 2019 and January 2022. Articles without the word "video" were dropped to increase the recall of retrieving video-form fake news. To summarize news events from abundant fact-checking articles, we designed heuristic regular expressions to extract key sentences and removed duplicate news events using K-means clustering based on sentence representations of BERT (Devlin et al. 2019). Then we paraphrased these key sentences in more general forms and finally obtained 854 event descriptions as search queries. Thereafter, we crawled relevant videos from two popular short video platforms[1] in China, *i.e.,* Douyin (the equivalent of TikTok in China) and Kuaishou. Besides video contents, we also crawled user responses and publisher profiles[2] as shown in Table 2.

[1]douyin.com, kuaishou.com

[2]Note that we only crawled the public information of the video's publisher for privacy reasons.

Data Annotation and Statistics

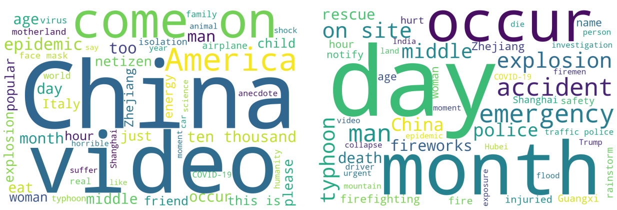
Although the searched queries are related to fake news that has been debunked, the retrieved videos are not always fake news videos. Therefore, we manually annotated 11,603 videos shorter than five minutes. The annotators were required to classify the given video into fake, real, debunked, and others (including useless and unverifiable videos). After balanced sampling, we ended up with 1,827 fake news videos, 1,827 real news videos, and 1,884 debunked videos under 738 events. Figure 3 shows the event and time distribution on the three classes of these videos. Additionally, the percentage of real and fake news videos with comments is 75% and 68% respectively.

Data Analysis

To reveal the different behaviors between fake and real news videos, we provide some exploratory analyses from three perspectives, *i.e.,* news content, social context and propagation, to offer insights for detection.

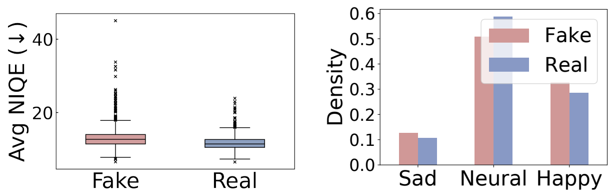
News Content

Text. Text is the dominant modality to detect fake news in the literature (Shu et al. 2017). We observe that fake news videos have shorter and more empty titles providing less information compared with real news. The word cloud of video titles in Figure 4 shows that fake news titles emphasize the word "video" much, prefer emotional and spoken words, and cover diverse topics. Differently, real news videos use more journalese and focus more on accidents and disasters.



(a) Fake (b) Real

Figure 4: Word cloud of video titles.



(a) Video frame quality. (b) Audio emotion.

Figure 5: Multimodal analysis.

Video. Image quality is considered to reflect originality and thus is used as an effective clue in text-image fake news detection (Qi et al. 2019; Cao et al. 2020). Following their work, we employ NIQE (Mittal, Soundararajan, and Bovik 2012) on video frames to indirectly measure video quality. Figure 5(a) shows that fake news videos have lower quality than real news and contain videos with particularly poor quality.

Audio. Emotion plays an important role in detecting fake news, and the emotional signals exist in the news text, news images or user responses (Zhang et al. 2021). Naturally, we analyze the speech emotion by the pre-trained wav2vec2 model (Baevski et al. 2020). Figure 5(b) shows that the speech in fake news videos shows more obvious emotional preferences than real news.

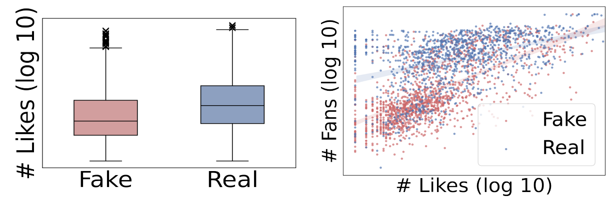
Social Context

Publisher Profiles. User profiles on social media have been shown to be correlated with fake news (Shu, Wang, and Liu 2018). As Figure 6(a) shows, most publishers of real news are verified accounts while most fake news publishers are not. Also, Figure 6(b) shows the average value of each user attribute after normalization. We could see that fake news publishers have more "consuming" behaviors (subscribes) and less "creating" behaviors (published videos, received likes, and fans) than real news publishers.

User Responses. We use the number of likes as an example to analyze user responses to fake and real news. Figure 7(a) shows that real news videos receive more likes than fake news, which is intuitive considering that real news publishers have more fans. To exclude the impact of account exposure, we analyze the relationship between the number of likes and the publisher's fans in Figure 7(b). We find that fake news videos receive more likes than real news when their publishers have a similar number of fans, which illustrates that fake news videos are more attractive than real news. In addition, the content of comments is also indicative in detecting fake news benefiting from crowd intelligence. According to our statistics, 18% fake news videos receive doubtful comments (*e.g.,* "Really?" and "Fake!") while the ratio is only 4% for real news.

(a) Authority (b) Statistics

Figure 6: Publisher's profile.



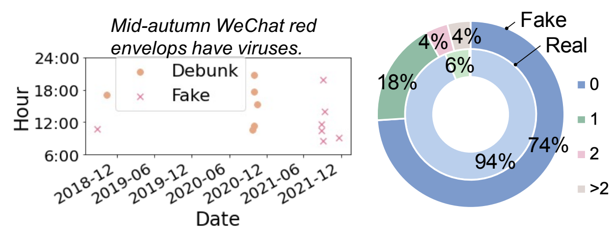
(a) Number of likes (b) Relationship between the number of publisher fans and likes.

Figure 7: User responses.

Propagation

Temporal Distribution. To study the effect of debunking videos in preventing the propagation of fake news, we visualize their published time and find that fake news that has been previously debunked can still spread; this further highlights the importance of automatic detection of fake news videos. According to our statistics, for 434 events with debunking videos, 39% of them have fake news videos emerged after the debunking videos were posted, especially the current or long-standing hot events. For example, as Figure 8(a) shows, the fake and debunking videos always occurred during the period of the mid-autumn festival (around September).

Video Duplication. Different from traditional social medias, there is no explicit propagation behaviors like retweeting on short video platforms. With the convenience of video editing functions provided by these platforms, people tend to edit and re-upload the videos, usually with no mention of the source. Therefore, we employ the pHash cluster algorithm (Klinger and Starkweather 2008) on the cover images to indirectly analyze the video duplication. As Figure 8(b)



(a) Temporal distribution in an event. (b) Video duplication.

Figure 8: Propagation analysis.

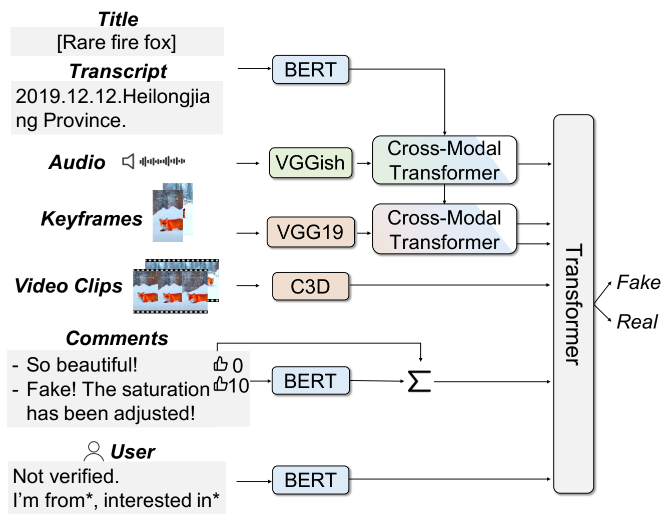


Figure 9: Architecture of the proposed framework SV-FEND.

shows, fake news videos have higher repetition while real news videos are more diverse. This phenomenon is due to the fact that real events will receive various images/videos taken by different witnesses (Jin et al. 2017b).

Method

Model Overview

Figure 9 illustrates the overall architecture of the proposed SV-FEND. In order to model the diverse characteristics of the input news video, we first extract features of multiple modalities, including the text, audio, keyframes, video clips, comments, and user. Next, we employ two cross-modal transformer layers to model the correlations between the text and the audio and keyframes. Another transformer layer is further used to dynamically fuse the multimodal features of news content and social context for the final classification.

Multimodal Feature Extraction

Title&Transcript. For video-form news, the embedded text in the video is usually more complete than the title of which the length is limited. Therefore, we extract the video transcript and concatenate it with the title. The composed text is further fed into the pre-trained BERT to extract the textual features $\boldsymbol{H}_T = [\boldsymbol{w}_1, ..., \boldsymbol{w}_l]$, where $\boldsymbol{w}_i$ represents the feature of the i-th word and l is the text length.

Audio. As a new imported modality compared to the text-image fake news, the audio modality includes speech, environment sound, and background music. Thus it provides not only semantics but also distinctive information in understanding the emotion of speakers and even the intent of the publisher. Specifically, we split the audio from the original video, and then use the pre-trained VGGish model (Hershey et al. 2017) to extract the audio features $\boldsymbol{H}_A = [\boldsymbol{a}_1, ..., \boldsymbol{a}_n]$, where n is the number of audio frames.

Video. We represent a video both at the frame and clip levels to obtain spatiotemporal and multi-granularity information for detection (Wu et al. 2015). Taking Fig. 1(a) as an example, tampered and untampered regions have inconsistent compression rates (frame level) and motion trajectories (clip level). At the frame level, we use the pre-trained VGG19 (Simonyan and Zisserman 2015) model to extract the static visual feature $\boldsymbol{H}_I = [\boldsymbol{i}_1, ..., \boldsymbol{i}_m]$, where m is the number of frames. We also take 16 subsequent frames centered at each time step as a video clip, and use the pre-trained C3D model (Tran et al. 2015) to extract the motion features $\boldsymbol{H}_V = [\boldsymbol{v}_1, ..., \boldsymbol{v}_m]$. An average operation is applied to obtain the aggregated motion feature $\boldsymbol{x}_V$.

Comments. Similar to text modality, we use the pre-trained BERT to extract the features of each comment, that is $\boldsymbol{H}_C = [\boldsymbol{c}_1, ..., \boldsymbol{c}_k]$, where k is the number of comments. Moreover, we use the number of likes to measure the importance of each comment. Specifically, the feature of comments is computed as:

$$\boldsymbol{x}_C = \sum_{j=1}^{k} \frac{l_j + 1}{\sum_{j=1}^{k} l_j + k} \boldsymbol{c}_j, \tag{1}$$

where l_j is the number of likes towards the j-th comments.

User. We concatenate the self-written and verified introduction of the publisher and feed it into the pre-trained BERT. The embedding of the $[CLS]$ token is extracted as the user features $\boldsymbol{x}_U$.

Multimodal Feature Fusion

The correlations between different modalities could help understand video semantics. For example, the acoustic characteristics of the audio modality often contain useful hints (such as adjusting the volume and tone) in capturing the key information in the speech text. The text could help select important ones from complex visual frames, and vice versa. Therefore, we use two cross-modal transformers to model the mutual enhancement between text and other modalities.

The first cross-modal transformer in Figure 9 takes the textual feature $\boldsymbol{H}_T$ and the audio feature $\boldsymbol{H}_A$ as input, and the audio-enhanced textual feature is further fed into the second cross-modal transformer to interact with the frame feature $\boldsymbol{H}_I$. Specifically, we use a two-stream co-attention transformer like (Lu et al. 2019) to process the multimodal information simultaneously, where the queries from each modality are passed to the other modality's multi-headed attention block. Finally, we obtain the textual feature enhanced by audio and visual frames $\boldsymbol{H}_{T \leftarrow A,I}$, text-enhanced audio feature $\boldsymbol{H}_{A \leftarrow T}$, and text-enhanced frame feature $\boldsymbol{H}_{I \leftarrow T}$. We average them and then obtain the features $\boldsymbol{x}_T, \boldsymbol{x}_A$, and $\boldsymbol{x}_I$.

Till now, we have extracted six features to represent the input video from different perspectives, and used two cross-modal transformers to model the correlations between different modalities in news content. Thereafter, we utilize self-attention (Vaswani et al. 2017) to model the correlations between features from news content and social context. Specifically, we concatenate the obtained six features into a feature sequence and feed it into a standard transformer layer, finally obtaining the multimodal fused feature $\boldsymbol{x}_m$.

Modality	Data	Method	Acc.	F1	Prec.	Recall
Social	M	hc feas+SVM	$67.98_{\pm1.12}$	$67.92_{\pm1.15}$	$68.11_{\pm1.08}$	$67.98_{\pm1.12}$
	C	hc feas+SVM	$64.27_{\pm2.08}$	$64.15_{\pm2.14}$	$64.63_{\pm2.27}$	$64.41_{\pm2.23}$
		BERT+Att	$62.87_{\pm3.52}$	$60.62_{\pm6.91}$	$63.33_{\pm4.57}$	$62.13_{\pm4.99}$
Text	T&Tr	hc feas+SVM	$70.22_{\pm2.10}$	$70.31_{\pm2.21}$	$70.37_{\pm2.18}$	$70.33_{\pm2.20}$
		Text-CNN	$74.66_{\pm2.31}$	$74.62_{\pm2.30}$	$74.79_{\pm2.41}$	$74.66_{\pm2.31}$
		BERT	$76.82_{\pm3.33}$	$76.80_{\pm3.34}$	$76.89_{\pm3.33}$	$76.82_{\pm3.33}$
Visual	F	VGG19+Att	$69.40_{\pm2.64}$	$69.33_{\pm2.57}$	$69.64_{\pm2.89}$	$69.40_{\pm2.64}$
		Faster R-CNN+Att	$70.69_{\pm2.80}$	$70.58_{\pm2.76}$	$71.03_{\pm2.99}$	$70.69_{\pm2.80}$
	V	C3D+Att	$69.05_{\pm1.71}$	$68.93_{\pm1.66}$	$69.36_{\pm1.87}$	$69.05_{\pm1.71}$
Audio	A	emo feas+SVM	$60.64_{\pm1.22}$	$60.61_{\pm1.24}$	$60.67_{\pm1.20}$	$60.64_{\pm1.22}$
		VGGish	$66.78_{\pm1.12}$	$66.63_{\pm1.13}$	$67.07_{\pm1.20}$	$66.78_{\pm1.12}$
Multimodal	M, Tr, A	(Hou et al. 2019)	$68.64_{\pm2.01}$	$68.01_{\pm1.99}$	$70.24_{\pm2.42}$	$68.64_{\pm2.01}$
	T, C	(Medina et al. 2020)	$71.45_{\pm2.43}$	$71.45_{\pm2.44}$	$71.47_{\pm2.42}$	$71.45_{\pm2.43}$
	T, C, F	(Choi and Ko 2021)	$75.07_{\pm0.38}$	$75.04_{\pm0.38}$	$75.18_{\pm0.38}$	$75.07_{\pm0.38}$
	Tr, F, A	(Shang et al. 2021)	$75.04_{\pm3.28}$	$75.02_{\pm3.29}$	$75.11_{\pm3.27}$	$75.04_{\pm3.28}$
	ALL	SV-FEND(ours)	**79.31**$_{\pm2.75}$	**79.24**$_{\pm2.79}$	**79.62**$_{\pm2.60}$	**79.31**$_{\pm2.75}$

Table 3: Performance (%) comparison between the proposed model and the baseline models. We report the mean and standard deviation of the five-fold cross-validation. Data formats are abbreviated for simplicity (M: Metadata, T: Title, Tr: Transcript, C: Comment, F: Keyframe, V: Video clip, A: Audio).

Classification

We use a fully connected layer with softmax activation to project the multimodal feature vector $\boldsymbol{x}_m$ into the target space of two classes: real and fake news videos, and gain the probability distributions:

$$\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{W}_b \boldsymbol{x}_m + \boldsymbol{b}_b), \tag{2}$$

where $\boldsymbol{p} = [p_0, p_1]$ is the predicted probability vector with p_0 and p_1 indicate the predicted probability of label being 0 (real news video) and 1 (fake news video) respectively. $\boldsymbol{W}_b$ is the weight matrix and $\boldsymbol{b}$ is the bias term. Thus, for each news post, the goal is to minimize the binary cross-entropy loss function as follows,

$$\mathcal{L}_p = -[(1-y)\log p_0 + y \log p_1], \tag{3}$$

where $y \in \{0, 1\}$ denotes the ground-truth label.

Experiments

Baseline Methods

To establish a comprehensive benchmark, we experiment with multiple representative methods that utilize single or multiple modalities on FakeSV.

Social Modality: For Metadata (**M**), we combine the **h**and-**c**rafted features (**hc feas**) of metadata proposed in (Hou et al. 2019; Li et al. 2022) and classify them by **S**upport **V**ector **M**achine (**hc feas+SVM**). For Comment (**C**), we use the **hc feas** about comments proposed in (Palod et al. 2019; Medina et al. 2020; Li et al. 2022) and **SVM** as classifier. Also, we use **BERT** to extract the features of multiple comments and use the **Attention** mechanism to fuse them for classification.

Text Modality-Title (**T**)&Transcript (**Tr**): We use **hc feas** about the title and transcript proposed in (Papadopoulou et al. 2018; Palod et al. 2019; Medina et al. 2020) and **SVM** as classifier. Also, we adopt **Text-CNN** (Kim 2014) and **BERT** as textual encoders with an MLP for classification.

Visual Modality: For keyframe (**F**), we use two visual encoders used in existing multimodal baselines, that is **VGG19** and **Faster R-CNN** (Ren et al. 2015), to extract the visual features of multiple frames. For video clip (**V**), we use the pre-trained **C3D** to extract the motion features. The **Attention** mechanism is used to fuse features of different time steps for classification.

Audio Modality (A): We extract the emotion audio features (**emo feas**) proposed in (Hou et al. 2019) and use **SVM** to classify. Also, we use the pre-trained **VGGish** model to extract the acoustic features for classification.

MultiModality: We use four existing state-of-the-art methods as multimodal baselines: 1) **(Hou et al. 2019)** use the linguistic features from the speech text, acoustic emotion features, and user engagement features and a linear kernel SVM to distinguish the real and fake news videos. 2) **(Medina et al. 2020)** extract tf-idf vectors from the title and the first hundred comments and use traditional machine learning classifiers including logistic regression and SVM. 3) **(Choi and Ko 2021)** use the topic distribution difference between title and comments to fuse them, and concat them with the visual features of keyframes. An adversarial neural network is used as an auxiliary task to extract topic-agnostic multimodal features for classification. 4) **(Shang et al. 2021)** use the extracted speech text to guide the learning of visual object features, use MFCC features to enhance the speech textual features, and then use a co-attention module to fuse the visual and speech information for classification.

Evaluation

Prior experiments have shown that the detection performance varies a lot under different data splits. To ensure fairness of experiments, our evaluations are conducted by doing five-fold cross-validation with accuracy (Acc.), macro precision (Prec.), macro recall (Recall), and macro F1-score (F1) as evaluation metrics. For each fold, the dataset is split as training and testing sets at the event level with a ratio of 4:1, ensuring that there is no event overlap among different sets.

Method	Acc.	F1	Prec.	Recall
SV-FEND	**79.31**	**79.24**	**79.62**	**79.31**
w/o News Content	74.89	74.57	76.34	74.89
w/o Text	75.37	75.11	76.54	75.37
w/o Title	77.28	77.23	77.53	77.28
w/o Transcript	77.89	77.86	78.01	77.89
w/o Visual	77.97	77.92	78.20	77.97
w/o Keyframes	78.49	78.43	78.77	78.49
w/o Video Clips	78.95	78.93	79.07	78.95
w/o Audio	78.95	78.93	79.03	78.95
w/o Social Context	78.62	78.54	78.98	78.62
w/o User	78.76	78.66	79.26	78.76
w/o Comment	79.09	79.02	79.45	79.09

Table 4: Ablation study on different modalities. The standard deviation values are ignored for simplicity.

Method	Acc.	F1	Prec.	Recall
(Hou et al. 2019)	71.89	71.29	73.88	71.89
(Medina et al. 2020)	75.58	75.50	75.92	75.58
(Choi and Ko 2021)	78.32	78.31	78.37	78.32
(Shang et al. 2021)	74.45	74.39	74.67	74.45
SV-FEND(ours)	**81.05**	**81.02**	**81.24**	**81.05**

Table 5: Performance comparison under temporal split.

Experimental Results

Performance Comparison. Table 3 shows the results of the baseline models and the proposed model, from which we can draw the following observations: 1) SV-FEND performs much better than the other methods, which validates that SV-FEND can effectively capture important multimodal clues to detect fake news videos. 2) According to the best-performing methods, the discriminability of different data could be sorted as: title&transcript >keyframe >video clip >metadata >audio >comments. 3) The best performance in terms of average accuracy on FakeSV is less than 0.8, which is lower than the accuracy achieved on other popular multimodal fake news datasets (Jin et al. 2017a; Boididou et al. 2016) in which the best accuracy is higher than 0.9. This further demonstrates the challenges of this dataset.

Analysis on Different Modalities. We conduct ablation experiments to analyze the importance of each modality in detecting fake news videos. From Table 4, we can see that 1) All modalities in SV-FEND are useful for achieving its best performance. 2) News content is more effective than social context in detecting fake news. 3) Different data under the same modality have homogeneity and complementarity, such as title and transcript, and keyframes and video clips. 4) Comments play the least role in the detection, which maybe results from data sparsity.

Performance on Temporal Split. In real-world scenarios, when a check-worthy news video emerges, we only have the previously-emerging data to train the detector. Therefore, we provide a temporal split in chronological order with a ratio of 70%:15%:15%, to evaluate the ability of models to detect future fake news videos. Table 5 validates the superiority of the proposed SV-FEND model in the temporal split. Also, we observe that the performance in the temporal split is higher than that in the event split, which is due to the existence of long-standing fake news events.

(a) Detected (b) Missed

Figure 10: Two representative fake news videos in FakeSV that were detected and missed by SV-FEND respectively.

Case Studies

To intuitively show the ability of SV-FEND, we conduct a qualitative analysis on those successful and failed examples. Figure 10(a) shows a successfully detected fake news video stitched together from two videos that have similar scenes but belong to different news events. Figure 10(b) shows a missed fake news video where the police drill was depicted as a real-life incident. It demonstrates the one-sidedness of videos uploaded by ordinary users and the limitations of the task of fake news detection without external information, which encourages us to further explore the combination of fake news detection and fact-checking on FakeSV.

Conclusion and Potential Applications

In this paper, we constructed the largest Chinese fake news short video dataset namely FakeSV. It contains abundant features about news content and social context, which not only provide a benchmark for different methods, but also support various researches related to fake news. The in-depth statistical analysis revealed the differences between fake and real news videos from multiple perspectives. Additionally, we provided a new multimodal baseline method and performed extensive experiments to compare the performance of different methods and modalities on FakeSV.

In addition to fake news video detection studied in this article, the multidimensional information provided in FakeSV can be useful for other applications as follows: 1) Detecting previously fact-checked fake news and propagation intervention. According to the phenomenon revealed in this article, fake news that has been previously fact-checked can still spread. Therefore, automatic detection of previously fact-checked fake news videos and personalized recommendations of debunked videos are indispensable. FakeSV provides debunking articles/videos corresponding to the fake news videos, which could support this task. 2) Fake news evolution analysis. By restoring the propagation (secondary editing) chain of videos based on video similarity, we could study the lifecycles of fake news events and the intentions of fake news publishers and spreaders. 3) User trustworthiness and susceptibility analysis. We could study the user trustworthiness by combining the published videos and the profile. By employing cross-platform analysis, we could compare users preferences and susceptibility to fake news across different platforms.

Ethical Statement

Following the literature that studied user profiles in fake news detection (Shu, Wang, and Liu 2018; Shu et al. 2019b; Dou et al. 2021), we collect the public information of the video publishers to study the behaviors of different user groups, *i.e.,* fake and real news publishers. We have anonymized the data and clearly stated what data is being collected and how it is being used in this article. Our data is meant for academic research purposes and should not be used outside of academic research contexts.

Acknowledgements

The authors would like to thank the annotators for their efforts. Also, we thank the anonymous reviewers for their valuable comments. This work was supported by the National Key Research and Development Program of China (2021AAA0140203), the National Natural Science Foundation of China (62203425), the Project of Chinese Academy of Sciences (E141020).

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, 12449–12460.

Boididou, C.; Papadopoulos, S.; Dang-Nguyen, D.-T.; Boato, G.; Riegler, M.; Middleton, S. E.; Petlund, A.; Kompatsiaris, Y.; et al. 2016. Verifying Multimedia Use at MediaEval 2016. In *Working Notes Proceedings of the MediaEval 2016 Workshop*.

Cao, J.; Qi, P.; Sheng, Q.; Yang, T.; Guo, J.; and Li, J. 2020. Exploring the Role of Visual Content in Fake News Detection. In *Disinformation, Misinformation, and Fake News in Social Media*, 141–161.

Choi, H.; and Ko, Y. 2021. Using Topic Modeling and Adversarial Neural Networks for Fake News Video Detection. In *30th ACM International Conference on Information & Knowledge Management*, 2950–2954.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Dou, Y.; Shu, K.; Xia, C.; Yu, P. S.; and Sun, L. 2021. User Preference-Aware Fake News Detection. In *44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2051–2055.

Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; et al. 2017. CNN Architectures for Large-scale Audio Classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 131–135.

Hou, R.; Pérez-Rosas, V.; Loeb, S.; and Mihalcea, R. 2019. Towards Automatic Detection of Misinformation in Online Medical Videos. In *2019 International Conference on Multimodal Interaction*, 235–243.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017a. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *25th ACM International Conference on Multimedia*, 795–816.

Jin, Z.; Cao, J.; Zhang, Y.; Zhou, J.; and Tian, Q. 2017b. Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Transactions on Multimedia*, 19(3): 598–608.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *2014 Conference on Empirical Methods in Natural Language Processing*, 1746–1751.

Klinger, E.; and Starkweather, D. 2008. pHash: The Open Source Perceptual Hash Library. http://www.phash.org/. Accessed June 14, 2022.

Li, X.; Xiao, X.; Li, J.; Hu, C.; Yao, J.; and Li, S. 2022. A CNN-based Misleading Video Detection Model. *Scientific Reports*, 12(1): 1–9.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *33rd International Conference on Neural Information Processing Systems*, 13–23.

Medina, S. J. C.; Papakyriakopoulos, O.; Hegelich, S.; and Hegelich, S. 2020. NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube. In *1st Workshop on NLP for COVID-19 at ACL 2020*.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3): 209–212.

Murayama, T. 2021. Dataset of Fake News Detection and Fact Verification: A Survey. *arXiv preprint arXiv:2111.03299*.

Palo, H. K.; Chandra, M.; and Mohanty, M. N. 2018. Recognition of Human Speech Emotion Using Variants of Mel-frequency Cepstral Coefficients. In *Advances in Systems, Control and Automation*, 491–498.

Palod, P.; Patwari, A.; Bahety, S.; Bagchi, S.; and Goyal, P. 2019. Misleading Metadata Detection on YouTube. In *European Conference on Information Retrieval*, 140–147.

Papadopoulou, O.; Zampoglou, M.; Papadopoulos, S.; and Kompatsiaris, I. 2018. A Corpus of Debunked and Verified User-Generated Videos. *Online Information Review*.

Qi, P.; Cao, J.; Li, X.; Liu, H.; Sheng, Q.; Mi, X.; He, Q.; Lv, Y.; Guo, C.; and Yu, Y. 2021. Improving Fake News Detection by Using an Entity-Enhanced Framework to Fuse Diverse Multimodal Clues. In *29th ACM International Conference on Multimedia*, 1212–1220.

Qi, P.; Cao, J.; Yang, T.; Guo, J.; and Li, J. 2019. Exploiting Multi-Domain Visual Information for Fake News Detection. In *19th IEEE International Conference on Data Mining*, 518–527.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, volume 28.

Shang, L.; Kou, Z.; Zhang, Y.; and Wang, D. 2021. A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. In *2021 IEEE International Conference on Big Data*, 899–908.

Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. 2019a. dEFEND: Explainable Fake News Detection. In *25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 395–405.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1): 22–36.

Shu, K.; Wang, S.; and Liu, H. 2018. Understanding User Profiles on Social Media for Fake News Detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval*, 430–435.

Shu, K.; Zhou, X.; Wang, S.; Zafarani, R.; and Liu, H. 2019b. The Role of User Profiles for Fake News Detection. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 436–439.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations*.

Sundar, S. S.; Molina, M. D.; and Cho, E. 2021. Seeing is believing: Is Video Modality More Powerful in Spreading Fake News via Online Messaging Apps? *Journal of Computer-Mediated Communication*, 26(6): 301–319.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning Spatiotemporal Features with 3d Convolutional Networks. In *IEEE International Conference on Computer Vision*, 4489–4497.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. volume 30.

Walker, M.; and Matsa, K. E. 2021. News Consumption Across Social Media in 2021. https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/. Accessed July 27, 2022.

Wu, Z.; Wang, X.; Jiang, Y.-G.; Ye, H.; and Xue, X. 2015. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *ACM Multimedia*, 461–470.

Zhang, X.; Cao, J.; Li, X.; Sheng, Q.; Zhong, L.; and Shu, K. 2021. Mining Dual Emotion for Fake News Detection. In *Web Conference 2021*, 3465–3476.