FakeSV: Multimodal Misinformation Detection in Short Videos

Team DeepTok
Justin Xiao, Eric Zhang, Zhiwei He, Zhangchi Fan

Introduction

Short video platforms like Douyin and Kuaishou have become major sources of news sharing — but also major breeding grounds for fake news. Existing fake news detection methods suffer from small datasets and insufficient use of multimodal and social context information.

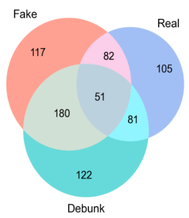
To address this, we re-implemented FakeSV, the largest Chinese fake news short video dataset, including video content, user comments, and publisher profiles.

SV-FEND is a multimodal detection model, which is the core component of FakeSV and is capable of:

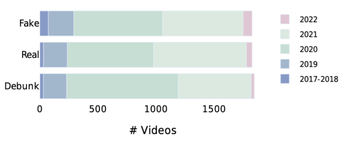
- Identify important features across different types of media like text, audio, and images.
- Combine content and social information (comments) to strengthen its detection.
- Accurately catches various kinds of fake news videos.

Dataset Info

Size:
1,827 fake news videos, 1,827 real news videos, and 1,884 debunked videos (total 5,538 videos) across 738 news events
Content:
•Video content (frames, audio, transcripts),
•Social context (user comments, likes, shares)



(a) Event



(b) Year

Methodology

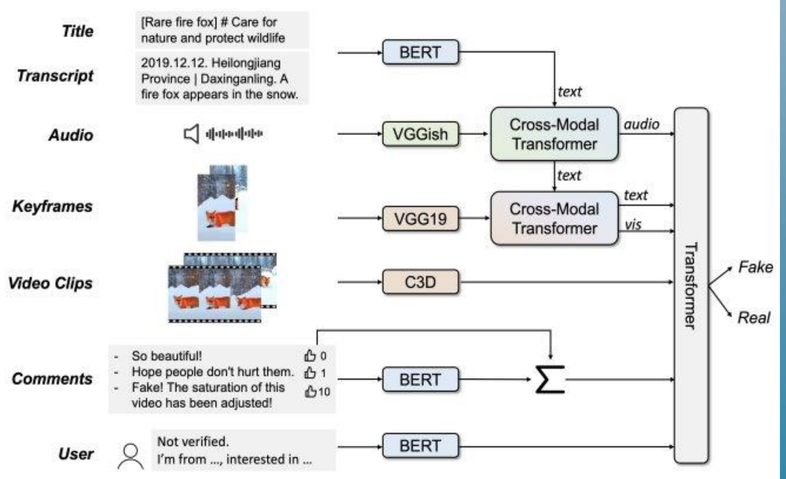
- **Multimodal Feature Extraction:**
  Extract features from video title + transcript (BERT), audio (VGGish), visual frames (VGG19), video clips (C3D), comments (BERT with like-weighted pooling).

- **Cross-modal Correlation Learning:**
  Two cross-modal transformers align and enhance textual, audio, and visual features.
  Helps spot key information across modalities (e.g., matching speech tone with video frames).

- **Social Context Fusion:**
  A self-attention transformer fuses multimodal content features with user comments and publisher information for final prediction.

- **Multimodal Feature Fusion and Representation:**
  Aggregate all enhanced features (text, audio, visual, comments, user profile) into a unified representation using a standard self-attention layer.

- **Classification:**
  Pass the final fused feature through a fully connected layer + Softmax to predict whether a video is real or fake, trained using binary cross-entropy loss.



Results

Due to training limitations, we trained on 1000 videos, including 517 fake news videos and 483 real news videos.

The results metrics (Acc, F1, Precision, Recall) compared with those in the paper are as follow in the table.

Metric	TF Re-implement	Original Pytorch SVFEND
Accuracy	72.50% ± 8.13%	79.31% ± 2.75%
F1-Score	63.79% ± 14.44%	79.24% ± 2.79%
Precision	82.60% ± 2.40%	79.62% ± 2.60%
Recall	66.52% ± 10.28%	79.31% ± 2.75%

Discussion

Issues:

- The FakeSV dataset is large and rich in multimodal information (videos, audios, comments, profiles).
- Some modules originally designed in PyTorch (e.g., Vggish) do not perform consistently when re-implemented in TensorFlow, leading to compatibility problems and unstable results.

Future Work:

- Adding more transformer layers and increasing model capacity to better capture complex multimodal interactions between text, audio, and visuals.
- Conduct ablation experiments to systematically evaluate the contribution of each modality (title, transcript, audio, visual frames, social context) to overall performance, guiding further model refinements.
- Implement data caching, memory mapping, and efficient batching strategies to speed up dataset reading and reduce memory usage during training.

Source

Qi, Peng, et al. "Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 12. 2023.

Qi, Peng. “ICTMCG/FakeSV: Official Repository for ‘Fakesv: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms’, AAAI 2023.” GitHub, github.com/ICTMCG/FakeSV/tree/main. Accessed 27 Apr. 2025.