MISINFORMATION DETECTION IN SOCIAL MEDIA VIDEO POSTS

Kehan Wang, David Chan, Seth Z. Zhao, John Canny, Avideh Zakhor

University of California, Berkeley
{wang.kehan, davidchan, sethzhao506, canny, avz}@berkeley.edu

ABSTRACT

With the growing adoption of short-form video by social media platforms, reducing the spread of misinformation through video posts has become a critical challenge for social media providers. In this paper, we develop methods to detect misinformation in social media posts, exploiting modalities such as video and text. Due to the lack of large-scale public data for misinformation detection in multi-modal datasets, we collect 160,000 video posts from Twitter, and leverage self-supervised learning to learn expressive representations of joint visual and textual data. In this work, we propose two new methods for detecting semantic inconsistencies within short-form social media video posts, based on contrastive learning and masked language modeling. We demonstrate that our new approaches outperform current state-of-the-art methods on both artificial data generated by random-swapping of positive samples and in the wild on a new manually-labeled test set for semantic misinformation.

***Index Terms*—** Multi-media forensics, misinformation detection, multimodal representation learning, deep learning for videos, social media

1. INTRODUCTION & RELATED WORK

Recent events, such as the COVID-19 pandemic and the 2020 US Presidential election have demonstrated that the spread of misinformation can cause relative chaos in times of uncertainty. Indeed, Vosoughi et al. [1] found in 2018 that the social media posts with falsified information spread faster, and reached more people than posts containing truthful facts. The emergence of short videos social media platforms, such as TikTok and Instagram, can additionally fuel the spread of misinformation.

To combat such misinformation, we develop methods to identify video posts that contain semantic inconsistencies, where a short video attached to the social media post does not semantically match its accompanying description. An example of semantic inconsistency is shown in Figure 1. The challenge of misinformation detection in social media video posts are two-folds: a) to learn a joint representation of video and text effectively; b) the lack of a large, labeled dataset for semantic matching. Here, we take steps towards both of these issues. To address the joint representation learning problem, we propose two deep-learning based methods for learning accurate multi-modal joint distributions, and utilize this representation to efficiently detect semantic inconsistencies. To address the data issues, we collect 160,000 social media video posts from Twitter to use as a large self-supervised training corpus, and introduce a novel testing dataset consisting of 401 professionally annotated videos to use as a gold standard for future unsupervised and self-supervised misinformation detection methods.

Pristine: COVID-19 vaccines are safe and effective. Get vaccinated today, with just one quick and painless dab!

Falsified: OMG look what they did to him!! Nobody should put chips in human! He can't breathe anymore!

Fig. 1: **Example of misinformation in a social media video post –** the video's mismatching text description in contains activity mismatch and topical shift.

There is a large body of literature on detecting multi-modal semantic inconsistencies. Luo et al. [2] leverage the expressiveness of a large pre-trained contrastive model CLIP [3] to classify misinformation based on retrieval. While methods based on billion parameter scale models can be powerful, many users do not have access to the compute or data required to train such models. Recently, several methods have been proposed for detecting differences in image and text semantics. Singhal et al. [4] leverages a learned joint embedding space, however requires both labeled positives and negatives in the data, and is specifically restricted to the news domain. Pan et al. [5] and Mayank et al. [6] focus explicitly on the textual description, detecting fake news using knowledge-graph based approaches. Tan et al. [7] and Fung et al. [8] focus on detecting synthetically generated news using text, image and knowledge element extraction. While these methods are feasible in situations where large labeled datasets of paired and unpaired semantic images and text exist, they do

not transfer well to the more complex and sparsely labeled video domain.

In the video/text domain, Shang et al. [9] use video, audio, text and metadata in TikTok videos to detect misleading COVID-19 video posts by fusing features from pre-trained models. Shang et al. does not, however, leverage the power of representation learning, and their method requires strong supervision, leading to generalization issues in low-resource domains. McCrae et al. [10] extract video, text, and named entity information from a news post, and utilize pretext-task learning on randomly permuted data to supervised a LSTM-based model. Unfortunately, because the method directly fuses video and text at each key-frame through concatenation, and does not learn a joint model of video and text, the model is unable to build complex joint representations.

In this paper, we introduce an extension of McCrae et al. [10], which solves the problem of superficial joint representations by making use of self-supervised representation learning in the form of both contrastive learning and masked language modeling to jointly model video and language.

2. METHODS

Our overall pipeline is shown in Figure 2a. Given a video post consisting of a video and a corresponding text description, we first use pre-trained models to extract video and text features. For text features, $s \in \mathbb{R}^{768}$, we use BERT [11], pre-trained on the masked language modeling task. For the video features, $v = (v_1, \dots v_n)$ $v_i \in \mathbb{R}^{512}$ we break the 10-fps video into segments of 32 frames, and use S3D [12] pre-trained on activity recognition, to extract one video feature per video segment. We explore 2 different methods in modeling joint video-language representation and detecting misinformation: Contrastive learning in Section 2.1 and Masked Language Modeling in Section 2.2.

2.1. Contrastive Learning (CL)

Our first method uses contrastive learning [13] to build the representation space of video and text, shown in Figure 2b. We first use a transformer encoder [14] to aggregate all information within the video features $v_i \in \mathbb{R}^{512}$. Using a Transformer allows long-range representation learning, rather than LSTMs, which suffer from numerous forgetting issues. We mean-pool the output from transformer encoder, $h_{1\dots n}$, to get one video feature v_{all}.

Given the embedded video feature $v_{all} \in \mathbb{R}^{512}$, and the text feature $s \in \mathbb{R}^{768}$, we use two projection layers to embed them onto the same feature dimension, $v'_{all}, s' \in \mathbb{R}^{P}$:

$$\begin{aligned} v'_{all} &= tanh(W_1 \cdot v_{all} + b_1) \\ s' &= tanh(W_2 \cdot s + b_2) \end{aligned} \tag{1}$$

where $W_1 \in \mathbb{R}^{512\times P}$, $W_2 \in \mathbb{R}^{768\times P}$, $b_1, b_2 \in \mathbb{R}^{P}$, and P is the projection dimension. After the projection, we use v'_{all} and s' as representations of the video and text features, respectively. At the projection dimension P, we use a cosine embedding loss, $L_{\cos}$ to construct the representation space of video and text:

$$L_{\cos}(v_{all}, s, y) = \begin{cases} 1 - cos(v_{all}, s) & \text{y=0} \\ max(0, cos(v_{all}, s) - margin) & \text{y=1} \end{cases} \tag{2}$$

Given two features v_{all} and s and their label y, 0 for match, 1 for mismatch, a cosine embedding loss $L_{\cos}$ encourages the cosine distance between matched samples to be smaller than the margin, and unmatched samples to be greater than the margin; see [15] for details. To perform misinformation detection, we concurrently concatenate v'_{all} and s' to obtain the joint representation $r \in \mathbb{R}^{2P}$, and use an MLP over the joint representation r to generate a likelihood of misinformation $l \in \mathbb{R}$, which we supervise with binary cross-entropy loss L_{BCE} shown below. Our final loss L_{all} is the mean of binary cross-entropy loss and cosine embedding loss:

$$\begin{aligned} r &= v'_{all} \oplus s' \\ l &= MLP(r) \\ L_{\text{BCE}} &= y \cdot log(\sigma(l)) + (1-y) \cdot log(1-\sigma(l)) \\ L_{\text{all}} &= 0.5 L_{\cos} + 0.5 L_{\text{BCE}} \end{aligned} \tag{3}$$

2.2. Masked Language Modeling (MLM)

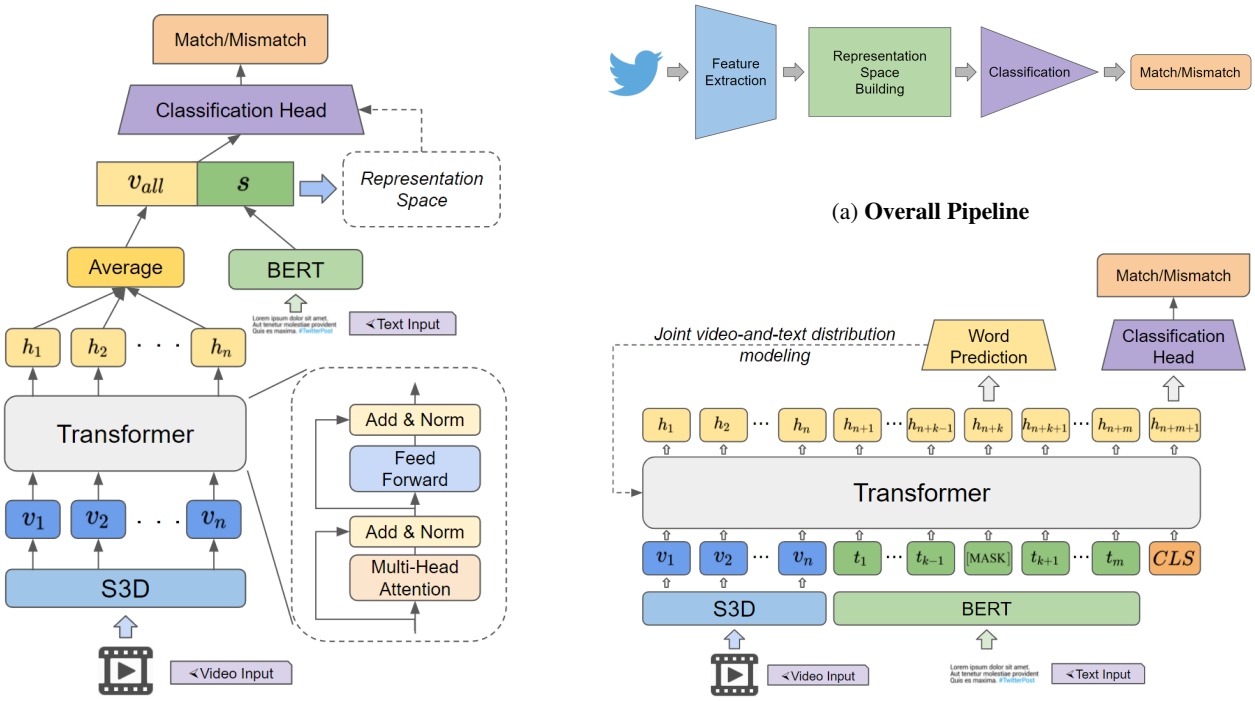
Our second method, shown in Figure 2c, models the joint distribution of video and text using a variation of Masked Language Modeling(MLM) proposed in BERT [11]. We train a transformer to approximate the maximum log-likelihood of each text token given its text context and the video,

$$E = \sum_{i}^{m} log(\mathbb{P}(t_i | t_{j\neq i}, v_{1..n}; \theta))$$

where $t_{1..m}$ are all m text tokens in video description, and θ represents parameters of the transformer, which are optimized through the masked language modeling objective from Devlin et al. [11].

To model the data, as in BERT[16], we use WordPiece [17] to tokenize each word of our text description with vocabulary size of 30522, and embed using a learned text embedding to obtain token embeddings $t_{1..m} \in \mathbb{R}^{768}$. We project our video features $v_{1..n}$ onto the same dimension $\mathbb{R}^{768}$ using a 2-layer MLP. We further append a learned classification token [CLS] $\in \mathbb{R}^{768}$ at the end of our sequence to extract all video-text information through encoding. Then, we randomly replace our text tokens with a special token [MASK], with a probability of 45% for each token. We construct our entire input embedding sequence as:

$$\begin{aligned} \text{video} &= (v_1, ..., v_n) \\ \text{masked\_text} &= (t_1, ...t_{k-1}, [\text{MASK}], t_{k+1}, ...t_m) \\ \text{input} &= \text{video} \oplus \text{masked\_text} \oplus [\text{CLS}] \end{aligned} \tag{4}$$



(a) **Overall Pipeline**

(b) **With Contrastive Learning**

(c) **With Masked-language Modeling (MLM)**

Fig. 2: Our overall pipeline and proposed methods in misinformation detection in a soical media video post.

Next, we add learned positional embeddings [14] to our input embedding sequence to capture the temporal order in video and text. We then apply a BERT-style transformer encoder with hidden dimension 768, feed-forward dimension 1024, and 12 layers on the input embedding sequence to receive hidden states $h_{1..(n+m+1)} \in \mathbb{R}^{768}$, which are finally projected onto the dimension of vocabulary size $\mathbb{R}^{30522}$. During training, we ask our model to reconstruct the original text tokens that were replaced, to learn each word's distribution within the context of the social media video post, $\mathbb{P}(t_i|\text{masked\_text}, v_{1..n})$. We use the cross-entropy reconstruction loss as our masked language modeling loss, L_{MLM}.

The last hidden state, h_{n+m+1}, of transformer output is the corresponding output of [CLS] token. We further apply a classification head on h_{n+m+1} and compute binary cross-entropy loss using the same method as in Section 2.1. Our final loss L_{all} is the mean of our masked language modeling loss and the binary cross-entropy loss:

$$L_{\text{all}} = 0.5L_{\text{MLM}} + 0.5L_{\text{BCE}} \tag{5}$$

3. EXPERIMENTAL DETAILS

Due to the lack of publicly available labeled dataset, we collect our own dataset using Twitter API. We scraped 160,000 tweets in English, with language labeling provided by Twitter, in the time frame of 2021. These tweets contain both a video ranging in length from 1 second to 10 minutes, with an average length of 44 seconds, and a short text description. To generate weakly supervised labels, we consider all videos and text descriptions of the 160,000 collected tweets as matching video and text pairs. By randomly swapping the text description of a video with another text description in the dataset, we create mismatching, semantic inconsistent video-and-text pairs. This random swapping procedure can produce misinformation that includes tonal/topical shifts, activity/object mismatches and other issues, however may also produce false-positives. The dataset is split into balanced train/validation/test divisions of 128k/16k/16k samples.

To compare with previous work, we fine-tune CLIP[3] on our training set using first frame of the video clips as its image input, as well as implement McCrae et al. 's [10] model without its Facebook post reactions input. We evaluate all methods by training and testing them on our random swapping dataset. As seen in Table 1, with explicit joint video-and-text modeling, both of our proposed methods outperform McCrae et al. [10] method by ∼8% and CLIP[3] by ∼35% on accuracy.

To measure how well the models perform against misinformation in the wild, we create a labeled test set of tweets. Four expert annotators were invited to label using video

Method	Accuracy	Precision	Recall
CLIP (ViT-B/32) [3]	59.24%	17.37%	**100.00%**
McCrae et al. [10]	85.83%	86.34%	85.30%
CL	94.33%	**94.12%**	94.34%
MLM	**94.51%**	92.73%	96.15%

Table 1: Performance on Random Swapping Dataset

Method	Accuracy	Precision	Recall
CLIP (ViT-B/32) [3]	23.44%	4.10%	**81.25%**
McCrae et al. [10]	62.84%	70.03%	80.43%
CL	65.84%	76.97%	79.22%
MLM	**71.07%**	**83.60%**	80.55%

Table 2: Performance on Manually Labeled Dataset

and text pairs sampled from the test division of our original 160,000 tweets. During labeling, a video and text pair is considered matched if the text description matches with the content of the video, and mismatched otherwise. The labeled test set contains 401 tweets, with 84 mismatched and 317 matched. All models' performance on this dataset is shown in Table 2. We see that Contrastive Learning outperforms [10]'s method by 3% on accuracy, and MLM performs the best overall, outperforming Contrastive Learning by 5.23% on accuracy. We speculate that the improvement in test accuracy in our model with MLM could be a result of (a) feeding all video and text tokens into the Transformer allows text tokens and videos to directly pay attention to each other to model their relationships better; and (b) compared with L_{cos} in contrastive learning, L_{MLM} makes the model more resilient to the dataset's bias, since its calculation does not rely on the random-swapping labels of match/mismatch. Therefore, our model using MLM is more robust to such a distribution shift from random swapping training dataset to a dataset of real-life misinformation.

We compare our proposed approaches with and without the representation space in Table 3. Models with representation space achieve higher accuracy in both datasets than models without it, supporting our key hypotheses. Noticeably, representation space improves our models' labeled dataset accuracy by more than 5%, suggesting that joint representation training is essential for in-the-wild performance.

Method	RS Accuracy	ML Accuracy
CL - no L_{cos}	93.51%	60.85%
MLM - no L_{MLM}	93.59%	65.33%
CL	94.33%	65.84%
MLM	94.51%	71.07%

Table 3: **Performance with/without representation space** – RS - Random-Swapping; ML - Manually Labeled.

(a) True Positive (b) True Negative (c) False Positive (d) False Negative

Fig. 3: Labeled dataset prediction successes and failures.

Figure 3 demonstrates the performance of our methods on some qualitative examples. Figure 3d shows a false negative. In this case, the video and text provided in this case are entirely unrelated, and the video could be representing higher-level symbolic representations, which remains a challenge for vision and language understanding. Figure 3c shows a false positive where the description focuses on tickets sale, while the video is showing a vote count. Likely, the numeric nature of the text and video were both inferred by the model, but the fine-grained semantics were not captured correctly. Generally, we found a number of the false positive cases to be weakly correlated video and text pairs. Future work involves learning representations that are sufficiently fine-grained to detect such mismatches.

4. CONCLUSION

In this work we have introduced two novel methods for joint video-and-text modeling designed to detect misinformation in social media video posts. Our new methods demonstrate significant improvements vs. state-of-the-art methods in both random-swapped and in-the-wild data. While leveraging self-supervised joint multi-modal representation learning has shown great improvement, we have also demonstrated that it still remains vulnerable to complex mismatches in real-wold misinformation. Future work involves developing higher-fidelity joint representations.

References

[1] Soroush Vosoughi, Deb Roy, and Sinan Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018. 1

[2] Grace Luo, Trevor Darrell, and Anna Rohrbach, "Newsclippings: Automatic generation of out-of-context multimodal media," *CoRR*, vol. abs/2104.05893, 2021. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," *CoRR*, vol. abs/2103.00020, 2021. 1, 3, 4

[4] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh, "Spotfake: A multi-modal framework for fake news detection," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 2019, pp. 39–47. 1

[5] Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu, "Content based fake news detection using knowledge graphs," in *SEMWEB*, 2018. 1

[6] Mohit Mayank, Shakshi Sharma, and Rajesh Sharma, "DEAP-FAKED: knowledge graph based approach for fake news detection," *CoRR*, vol. abs/2107.10648, 2021. 1

[7] Reuben Tan, Bryan Plummer, and Kate Saenko, "Detecting cross-modal inconsistency to defend against neural fake news," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Nov. 2020, pp. 2081–2106, Association for Computational Linguistics. 1

[8] Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil, "InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection," 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)*, Online, Aug. 2021, pp. 1683–1698, Association for Computational Linguistics. 1

[9] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang, "A multimodal misinformation detector for covid-19 short videos on tiktok," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 899–908. 2

[10] Scott McCrae, Kehan Wang, and Avideh Zakhor, "Multi-modal semantic inconsistency detection in social media news posts," 2021. 2, 3, 4

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. 2

[12] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman, "End-to-End Learning of Visual Representations from Uncurated Instructional Videos," in *CVPR*, 2020. 2

[13] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 539–546 vol. 1. 2

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," 2017. 2, 3

[15] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman, "Self-supervised multimodal versatile networks.," *NeurIPS*, vol. 2, no. 6, pp. 7, 2020. 2

[16] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, "Is space-time attention all you need for video understanding?," *arXiv preprint arXiv:2102.05095*, 2021. 2

[17] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016. 2