Multi-Modal Fine-Grained Fake News Detection with Dialogue Summarization

Faiaz Rahman
CS 677: Advanced Natural Language Processing
Department of Computer Science
Yale University
faiaz.rahman@yale.edu

Abstract

The proliferation of fake news online has created an era of both digital misinformation and public mistrust, particularly on social media platforms where users can engage in dialogue about such content. Fake news exists not only in text form, but also includes any accompanying images and video with the original post. Prior work has begun exploring multi-modal models (e.g., consisting of multiple modes of input like text and images) for fake news detection by both developing datasets for experimentation and by examining different multi-modal representations. Given that user dialogue (e.g., comment threads, tweet replies, etc.) can often give more insight into the integrity of a post (e.g., by indicating how extreme of a response was garnered, etc.), we extend existing research on multi-modal models for fake news detection by integrating joint textual-visual embeddings with dialogue (i.e. comment thread) data to see if this can further improve performance. Our models incorporating text, image, and dialogue components outperform those with only text and image components by 14.1%, 9.5%, and 9.3% when classifying fake news in 6, 3, and 2 ways, respectively, illustrating the value of incorporating user dialogue data when detecting fake news on online platforms.[1]

1 Introduction

The proliferation of fake news online has created an era of both digital misinformation and public mistrust, particularly on social media platforms where users can engage in dialogue about such content. According to the 2021 Pew Research Center survey, 86% of U.S. adults get their news from a digital device (e.g., a smartphone, computer, or tablet) and 53% of U.S. adults get news from social media. As a result of its provocative design, fake news spreads more frequently and more quickly when compared to true, fact-based news (Vosoughi et al., 2018).

Fake news exists not only in text form, but also includes any accompanying images and video with the original post. We define models which incorporate multiple modes of data (e.g., text and images) as *multi-modal models*, and explore the power of such models in fake news detection in a more *fine-grained* approach, in which we detect if a post is fake news, and if so, its specific type (from satire, false connection, imposter content, manipulated content, and misleading content).

Prior work has begun exploring multi-modal models (e.g., consisting of multiple modes of input like text and images) for fake news detection by both developing datasets for experimentation and by examining different multi-modal representations. However, an area with less research is that which explores if *user dialogue* in response to a post (e.g., comment threads, tweet replies, etc.) can act as an additional mode of input to a model, and in turn help improve performance in fine-grained fake news detection. Given that user dialogue (e.g., comment threads, tweet replies, etc.) can often give more insight into the integrity of a post (e.g., by indicating how extreme of a response was garnered, etc.), we extend existing research on multi-modal models for fake news detection by integrating joint textual-visual embeddings with dialogue (i.e. comment thread) data.

2 Related Work

2.1 Fake News Detection

Pérez-Rosas et al. (2018) achieve accuracies of up to 76% using for binary fake news classification when using linear SVM classifiers on data annotated with linguistic features. More recently,

[1]For reproducibility and extensibility in future work, we make our code, which includes an automated process for running experiment configurations, available at https://github.com/faiazrahman/Multimodal-Fake-News-Detection.

Hansen et al. (2021) explore the process of models judging fake news based on their reasoning, i.e. given a claim with associated evidence, the model assesses the veracity of the claim based on the evidence, in the domain of political fact-checking. Karimi et al. (2018) introduce a framework for multi-source multi-class (i.e., fine-grained) fake news detection, where there multiple sources include a statement, metadata, history, and a report; we note that although these are indeed multi-source, they are not multi-modal. More recently, Wang et al. (2021) implement an approach using meta-learning to detect fake news on newly-emerging events with few verified posts.

There are many fake news datasets available (Wang, 2017; Shu et al., 2019; Thorne et al., 2018); however, a majority of the datasets consist of only text-based data. Fakeddit (Nakamura et al., 2020), the dataset we use, is a large multi-modal dataset with text, image, and dialogue data with multiple k-way classification labels.

2.2 Dialogue Summarization

Automatic text summarization aims to output the most salient parts of a given corpus, either in an extractive manner (where salient portions of the given corpus are extracted and pieced together into a summary) or an abstractive manner (where the model generates text of its own, e.g., as if it were writing "in its own words"). Dialogue summarization, or conversation summarization, aims to summarize a set of utterances. Gliwa et al. (2019) introduced SamSum, a chat-dialogue dataset for the conversation summarization task. Fabbri et al. (2021) introduced ConvoSumm, a dataset covering four dialogue domains of news article comments, discussion forums, community question-answering, and email threads.

3 Approach

3.1 Model Architecture

The main bulk of our overall model architecture are the three feature encoding models used to encode the text, image, and dialogue features; these embeddings are then combined into a multi-modal embedding, which is then passed into a feedforward network for fine-grained classification.

3.1.1 Multi-Modal Feature Encoding

Each post (i.e., each element in the dataset) consists of text, image, and dialogue components. We describe the encoding process for each component feature in the following.

Text Features The raw input text is passed through a Transformer-based encoder to get an embedding; we use an approach which takes variable-length input text and produces a fixed-size text embedding. We experiment with both RoBERTa and MPNet as our text encoder models.

Let v_t^* denote variable-length raw text size and d_t the text embedding dimension.

$$\texttt{EncodeText}: \quad \mathbb{R}^{v_t^*} \mapsto \mathbb{R}^{d_t}$$

Image Features The raw image input is first preprocessed into RGB channels, then resized and normalized to fixed dimensions. The preprocessed image tensor is then passed to ResNet (specifically, ResNet-152). The last layer of ResNet is overwritten with a fully-connected linear layer to get the image features (rather than a classification).

Let v_i^* denote variable-length raw image size, d_n represent the resized and normalized image dimension for one channel (i.e. RGB), and d_i the image embedding dimension.

$$\texttt{EncodeImage}: \quad \mathbb{R}^{v_i^*} \mapsto \mathbb{R}^{d_n \times d_n \times d_n} \mapsto \mathbb{R}^{d_i}$$

Dialogue Features and Summarization Pipeline The raw dialogue consists of a set of comments associated with each post. We utilize abstractive dialogue summarization to build our dialogue feature embedding, noting that many other alternative approaches can be explored in future work. For our dialogue summarization approach, we sort the comments by the number of up-votes they receive (which is included in the metadata) in descending order and pass the sorted sequence into a Transformers-based summarization pipeline; for our experiments, we use a BART summarization pipeline. We also note that some posts have a large number of comments, which in turn exceeded the input token limit of the summarization pipeline. In that case, we simply truncate comments beyond that limit, and given our sorting method, those are lower-value comments (as measured by the upvotes metadata).

The generated output summary text is then passed through a Transformer-based encoder to get the final dialogue summary embedding; we again experiment with both RoBERTa and MPNet as our text encoder models.

Let v_d^* denote variable-length raw dialogue utterance size, d_d the dialogue embedding dimension.

$$\texttt{EncodeDialogue:} \quad [\mathbb{R}^{v_d^*}, ...] \mapsto \mathbb{R}^{d_d}$$

We also note that when generating our summaries, we determine our generated summary's minimum and maximum length bounds by the following heuristic (computed sequentially as shown). This was found empirically to yield good summary quality while avoiding implementation-specific issues of the Transformers pipeline (e.g., a maximum length of 1 yields unexpected behavior, hence the second update of the maximum length variable). This heuristic could be iterated on in future work to further improve summary quality.

$$n = \sum_{w \in \text{ utterance}} 1$$
$$\text{max\_length} = \min\left(75, \frac{n}{2}\right)$$
$$\text{max\_length} = \max(\text{max\_length}, 5)$$
$$\text{min\_length} = \min(5, \text{max\_length} - 1)$$

3.1.2 Multi-Modal Embedding

Once the individual embeddings are obtained, they are combined through tensor concatenation followed by a feedforward fully-connected layer (with dropout); the output of the feedforward layer is the final multimodal embedding, which is then used in the classification pipeline.

Let d_{MME} denote the final multi-modal embedding dimension.

$$\texttt{MultimodalEmbed:}$$
$$\mathbb{R}^{d_t} \times \mathbb{R}^{d_i} \times \mathbb{R}^{d_d} \mapsto \mathbb{R}^{d_t+d_i+d_d} \mapsto \mathbb{R}^{d_{\text{MME}}}$$

3.1.3 Fine-Grained Classification

The multi-modal embedding is then passed through two fully-connected feedforward layers, i.e., mapping from the embedding to a hidden layer, and then to the final output layer with k nodes, where k is the k-way classification being done. (In our experiments, $k \in [6, 3, 2]$.)

Let d_h denote the hidden dimension.

$$\texttt{Classify:} \quad \mathbb{R}^{d_{\text{MME}}} \mapsto \mathbb{R}^{d_h} \mapsto \mathbb{R}^k$$

Thus, the overall model architecture allows for modular embedding of the individual modalities prior to their fusion; the resulting multi-modal embedding is then used in the classification portion of the overall model to yield the fine-grained prediction.

4 Data

Fakeddit (Nakamura et al., 2020) is a multi-modal dataset consisting of over 1 million samples from multiple categories of fake news, labeled with 2-way, 3-way, and 6-way classification categories to allow for both binary classification and, more interestingly, fine-grained classification. The dataset was collected from a diverse array of topic categories (i.e., subreddits) from Reddit, and includes a post's text, image, and comment threads. Given our compute resources and timing constraints with the array of experiments we were running, we used a randomly-sampled subset of their train and test datasets with a balanced class distribution (and selecting only examples which were multi-modal), consisting of 10,000 training examples and 1,000 evaluation examples. Future work can take advantage of our hyperparameter tuning and experimental settings to run experiments on larger subsets of the dataset.

Nakamura et al. (2020) ran experiments comparing multi-modal input of text and image data with single-modal input of text data and image data individually, finding that the multi-modal approach of using text and images simultaeously improved performance. However, despite collecting the comment thread data, the authors did not run any experiments integrating the comment thread data into their input representations, instead leaving the comment data (and additional metadata) for future work, which we pick up from.

The Fakeddit dataset is made publicly available[2] for usage; our codebase has a slightly modified version of their image downloading script.

5 Experiments

We run experiments to compare (1) the performance of text-image multi-modal models with text-image-dialogue multi-modal models, and (2) the performance of different text encoder models. We do not compare with single-modal models, since Nakamura et al. (2020) already compared text-image multi-modal models with single-modal text and single-modal image models and found that the multi-modal approach indeed had better performance. Thus, we focus on quantifying the performance of including dialogue data via dialogue summarization.

[2] https://github.com/entitize/Fakeddit

Modality	Models	Results		
		2-way	3-way	6-way
Text + Image	RoBERTa + ResNet	0.8035	0.8102	0.7078
	MPNet + ResNet	0.8212	0.8107	0.7250
Text + Image + Dialogue	RoBERTa + ResNet + BART	0.8870	0.8928	0.8237
	MPNet + ResNet + BART	**0.9141**	**0.9052**	**0.8669**

Table 1: Results from experiments. Note that the models refer to the text encoder model, the image encoder model, and the dialogue summarization model, respectively.

Measure	Models	Performance Improvement		
		2-way	3-way	6-way
Absolute Accuracy	RoBERTa + ResNet + BART	+8.35%	+8.26%	+11.59%
	MPNet + ResNet + BART	**+9.29%**	**+9.45%**	**+14.19%**
Relative Accuracy	RoBERTa + ResNet + BART	+10.17%	+10.20%	+16.37%
	MPNet + ResNet + BART	**+11.30%**	**+11.66%**	**+19.57%**

Table 2: Performance improvements measured in terms of absolute accuracy and relative accuracy when using multi-modal text–image–dialogue models (i.e. by including dialogue data via dialogue summarization), when compared to the baseline of using text–image models.

5.1 Models

For our text encoder, we experiment with both RoBERTa and MPNet. Specifically, our implementation uses `all-distilroberta-v1` and `all-mpnet-base-v2` from `sentence-transformers`.

For our image module, we use ResNet-152 via `torchvision.models.resnet152`.

For our dialogue summarization, we use a BART summarization pipeline from HuggingFace's `transformers`, specifically with the `sshleifer/distilbart-cnn-12-6` model. For encoding the generated dialogue summary, we use the same models as the text encoder (i.e., RoBERTa and MPNet).

5.2 Experiment Settings

We train on two NVIDIA K80 GPUs with Driver version 465.19.01 and CUDA version 11.3. We use PyTorch (version 1.10.0) to implement our models, including using PyTorch Lightning for our model training and evaluation. We run training and evaluation on both GPUs in data parallel (i.e., splitting each batch across both GPUs), with a batch size of 32 (and thus each GPU processing 16 items per batch, with the root node aggregating the results). We use a learning rate of `1e-4`, dropout percentage of 0.1, and train for 5 epochs. We use Adam as our optimizer and cross-entropy as our loss function.

We ran hyperparameter tuning and found that learning rates of `1e-3` and `1e-5` caused the loss to not decrease. We also found that using SGD as our optimizer (both with and without momentum) caused the loss to not decrease. We hypothesize that this was likely due to the algorithm converging early by getting stuck in a local minima; in the case of the learning rate, our final value of `1e-4` served as a good middle ground, and in terms of the optimizer, Adam simply outperformed SGD.

6 Results and Analysis

The results of our experiments are recorded in Table 1. When comparing the text encoder models of RoBERTa and MPNet, we find that MPNet slightly outperforms RoBERTa in all experiment configurations (i.e., for both the text-image modality and the text-image-dialogue modality, in all k-way classifications).

More signficantly, we find that the incorporation of dialogue data through dialogue summarization into our multi-modal models does indeed improve performance. When comparing within our RoBERTa-ResNet architecture, adding the dialogue modality increases absolute accuracy by 8.35%, 8.26%, and 11.59% for 2-, 3-, and 6-way fake news classification, respectively. (These are relative accuracy increases of 10.17%, 10.20%, and 16.37%, respectively.) When comparing within our MPNet-ResNet architecture, adding the dialogue modality increases absolute accuracy by 9.29%, 9.45%, and 14.19% for 2-, 3-, and 6-way fake news classification, respectively. (These are relative ac-

curacy increases of 11.30%, 11.66%, and 19.57%, respectively. These performance improvements are recorded in table 2.

It is interesting to note that finer-grained fake news detection tasks received larger performance improvements; in other words, the 6-way fake news classification task had the largest performance improvement when using a multi-modal modal that incorporated dialogue data via dialogue summarization. On the other hand, 2-way fake news classification had the least amount of a performance improvement. We hypothesize that this is because the 6-way fake news detection task is the most difficult, given that it involves not only detecting fake news in a binary fashion but also distinguishing between the five types of fake news (satire, false connection, imposter content, manipulated content, and misleading content, as defined in Section 1); as a result, it benefits the most from having additional information.

It is also interesting to note that incorporating dialogue data could have been a potential source of noise for the overall model, particularly since user dialogue online (e.g. comment threads, tweet replies) is often not highly-moderated and could essentially contain any text. We hypothesize that our method of sorting the comments by upvotes (and more generally, by using metadata available with individual dialogue utterances) and truncating comments which exceeded the BART Transformer summarization pipeline's input token length helped to both emphasize the high-value comments as important (based on their location at the beginning of the corpus prior to summarization) and remove low-value comments (in the case in which there were many comments and thus truncating was required).

7 Conclusion

Fake news exists online in a multi-modal fashion, yet we can leverage the user dialogue aspect of social media platforms and other online sites to further improve multi-modal fine-grained fake news detection. Our experiments yielded promising results, illustrating significant performance improvements when using multi-modal text–image–dialogue models (incorporating dialogue data via dialogue summarization) in comparison to only text–image models.

Further work can be done experimenting with both different approaches for summarization (e.g., different models, abstractive vs. extractive summarization) and different ways to incorporate the dialogue data (e.g., graph-based approaches). Future work can also explore textual-visual similarity (e.g., jointly training a network to learn the similarity between a post's text and its corresponding image, similar to that of Zhou et al. (2020), which is particularly relevant in the case of "clickbait"-style fake news).

We hope that our work can continue to drive forward the importance of building ethical natural language processing systems to mitigate the spread of misinformation in the ever-growing, ever-ubiquitous digital world.

References

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. 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 6866–6880, Online. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Casper Hansen, Christian Hansen, and Lucas Chaves Lima. 2021. Automatic fake news detection: Are models learning to reason?

Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. 2018. Multi-source multi-class fake news detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1546–1557, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection.

Yaqing Wang, Fenglong Ma, Haoyu Wang, Kishlay Jha, and Jing Gao. 2021. Multimodal emergent fake news detection via meta neural process networks. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: Similarity-aware multi-modal fake news detection.