Unimodal Ensemble: Addressing Modality Laziness in Multimodal Fake News Detection

David Gao
Department of Computer Science
Vanderbilt University
Nashville, United States
david.gao@vanderbilt.edu

Lincoln Murr
Department of Computer Science
Vanderbilt University
Nashville, United States
lincoln.d.murr@vanderbilt.edu

Haoli Yin
Department of Computer Science
Vanderbilt University
Nashville, United States
haoli.yin@vanderbilt.edu

Abstract—**The growth of fake news and misinformation on social media platforms has become a significant challenge in the digital age. Although there are existing multimodal fake news detection approaches, they often suffer from modality laziness, where the model primarily learns from a dominant modality, leading to suboptimal performance. We propose a novel Unimodal Ensemble (UME) architecture that independently trains unimodal image and text models in parallel to address this issue. By allowing each modality to learn task-relevant features without cross-modal influence, our method ensures that textual and visual information are effectively used in the fake news detection task. We benchmark our approach on two widely used datasets, Fakeddit and Weibo, and compare UME's performance against state-of-the-art models. Our evaluation demonstrates that our model outperforms these benchmarks, exhibiting competitive or state-of-the-art performance in multimodal fake news detection for various metrics. Our ablation study further validates the effectiveness of the UME approach compared to past proposed solutions for modality laziness. The code can be found at https://github.com/Nano1337/ume-fakenews.**

Index Terms—**Fake News Detection, Multimodal, Modality Laziness, Ensemble Learning**

I. Introduction

A. Motivation

In the era of rapid information growth and the proliferation of generative AI tools, spreading misinformation has become a significant challenge [22]. The internet, while serving as an invaluable resource for learning, has also become a breeding ground for fake news [8]. This issue is particularly concerning in educational settings, where students often explore new subjects and may lack the experience to discern between credible sources and fabricated content. It is also rampant across social media platforms, a medium that can potentially influence many real-world events, such as elections.

The impact of misinformation on the real world cannot be overstated [12]. When students are exposed to fake news, it can lead to misconceptions, hinder critical thinking skills, and ultimately undermine the educational process [21]. In the context of political campaigns, where technology is increasingly integrated into the voter experience, addressing the threat of fake news becomes even more crucial [2]. As such, there is a pressing need for effective fake news detection systems to support people in their quest for knowledge and help foster a culture of information literacy.

B. Challenges

Developing a fake news detection model presents several challenges. One major hurdle is overcoming bias in the data and ensuring the model can generalize well to various topics and information sources encountered in educational contexts. Information comes in many shapes and forms, so we must find datasets that encapsulate various scenarios.

Another significant challenge lies in contextualizing the information being analyzed. Fake news often relies on subtle manipulations of language and imagery that can be difficult to detect without considering the broader context. Models must understand the relationships between different information modalities and reason about their veracity holistically. This requires sophisticated techniques that can capture the semantic and contextual cues present in the data.

Furthermore, the accessibility and interpretability of fake news detection models are critical factors in their successful adoption in educational settings. Students and educators must be able to interact with the system and understand its outputs easily. This necessitates the development of intuitive interfaces and clear explanations of the model's decision-making process. Striking the right balance between model complexity and user-friendliness is a key challenge that must be addressed.

On a technical level, multimodal learning suffers from the ”modality laziness” problem [29], which prevents the fusion-based multimodal model from effectively learning from multimodal data, leading to possible underperformance, even when compared to its unimodal counterpart. We aim to more carefully explore how this phenomenon affects modeling multimodal fake news detection and if a solution exists to bypass this issue.

C. Novelty

Our proposed solution aims to tackle the challenges of fake news detection through a novel application of multimodal learning techniques. By leveraging textual and visual information, we seek a more comprehensive understanding of the analyzed content. Integrating multiple data modalities allows our model to identify subtle inconsistencies and manipulations that single-modality approaches may miss for greater robustness.

Central to our solution is a specific data fusion process tailored to the unique requirements of fake news detection. We explore techniques such as late fusion, on-the-fly gradient modulation with generalization enhancement (OGM-GE) [15], quality-aware multimodal fusion (QMF) [28], and unimodal model ensembling (UME), to determine the most effective approach for combining the outputs of the text and image models (see Section VI.B for benchmarking). This comparative study of fusion techniques sets our work apart from previous efforts and contributes to the broader understanding of multimodal learning in the context of fake news detection.

D. Contributions

By addressing the challenges of fake news detection in a real-world context through a novel multimodal approach, our work aims to empower people with the tools and knowledge necessary to navigate the complexities of the modern information landscape. We believe that our contributions will enhance the learning experience and foster a culture of critical thinking and information literacy that extends to more applications in the future.

II. Related Work

Fake news and misinformation detection have garnered significant attention recently due to the increasing prevalence of visual and textual disinformation on social media platforms. Researchers have proposed various approaches for tackling the problem, focusing on textual and visual modalities.

Several works emphasize improving the detection of out-of-context captions in multimodal fake news. [1] builds upon the COSMOS framework [3] and proposes four methods to improve its detection accuracy, including fake-or-fact checking, differential sensing, object-caption matching, and threshold adjustment. Similarly, [9] incorporates textual semantics understanding from large corpora and combines it with various combinations of text-image matching and image captioning methods to classify triples of (image, caption1, caption2) into out-of-context and no out-of-context labels.

Another research gap focuses on the data-intensive nature of deep neural networks and the need to preserve structural and semantic information in multimodal fake news detection. [30] introduces a self-supervised model grounded in contrastive learning to extract features from text and image simultaneously, achieving strong visual feature extraction with minimized data training requirements. Liu et al. [11] propose a semantic distillation and structural alignment (SDSA) network to reduce redundant information and effectively preserve structural and semantic information.

”Synthetic Misinformers: Generating and Combating Multimodal Misinformation” [14] explores manners by which we can automatically generate multimodal misinformation synthetically to tackle the labor-intensive process of manual annotation. The authors comparatively study existing and new Synthetic Misinformers and demonstrate that CLIP-based Named Entity Swapping can lead to better multimodal detection models.

[10] introduces a consistency-learning fine-grained fusion network (CFFN) that explores the consistency and inconsistency from high and low relevant word-region pairs. The authors employ a cross-modal attention mechanism to evaluate consistency and calculate inconsistency scores, demonstrating CFFN's superior performance compared to baselines.

There are four papers representing the current state of the art.

Wang et al. [23] introduce an end-to-end framework that learns event-invariant features enabling fake news detection on newly-emerged events. Their Event Adversarial Neural Network (EANN) includes a multimodal feature extractor, a fake news detector, and an event discriminator. The feature extractor and fake news detector together learn discriminable representations for fake news detection, while the event discriminator removes event-specific features to retain shared features among events.

SpotFake [20] introduces a multimodal approach for fake news detection that addresses the issue of relying on subtasks like event discriminators or cross-modality correlations. The authors exploit textual and visual features of an article without considering other subtasks. The BERT language model is used to learn text features, and VGG-19, pre-trained on ImageNet, learns image features. It performs highly on the Twitter and Weibo datasets. In a follow-up work, SpotFake+ [19] extends SpotFake by utilizing transfer learning to capture semantic and contextual information from news articles and associated images. It is the first work to perform a multimodal approach for fake news detection on a dataset with full-length articles.

In [16], authors propose a hierarchical multimodal contextual attention network (HMCAN) that can jointly model multimodal context information and the hierarchical semantics of text in a unified deep model. HMCAN employs BERT and ResNet to learn better representations for both text and images, which is then used as data in a multimodal contextual attention network to combine inter-modality and intra-modality relationships. Fake news detection is achieved using a hierarchical encoding network that captures hierarchical semantics.

Lastly, ”Cross-Modal Ambiguity Learning for Multimodal Fake News Detection” [5] addresses ambiguity across different modalities that leads to inferior detection. CAFE, an ambiguity-aware multimodal fake news detection method, consists of a cross-modal alignment module, a cross-modal ambiguity learning module, and a cross-modal fusion model. CAFE aggregates unimodal features and cross-modal correlations adaptively based on the strength of the cross-modal ambiguity, which leads to more accurate fake news detection.

Our proposed solution differs from the aforementioned works in several key aspects. Despite the advancements made by these papers, a key weakness they share is the problem of modality laziness. This phenomenon occurs in multimodal fusion models (especially late-fusion models) where the multimodal model underperforms its unimodal counterpart by primarily learning from the dominant modality that emerges from gradient-based training dynamics that's faster/easier to learn [29]. This in turn saturates the joint multimodal loss (e.g.

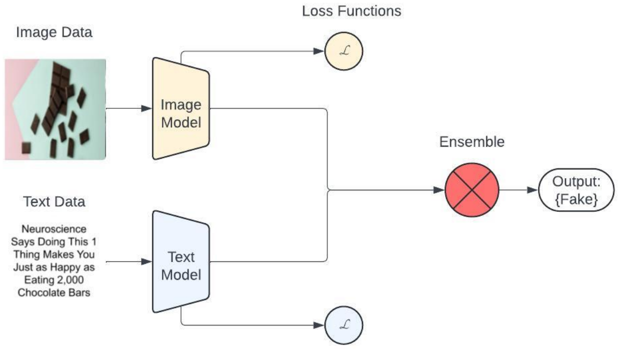

Fig. 1. Unimodal Ensemble (UME) Architecture with unimodal loss functions. The ensemble operation averages the unimodal logits to attain a joint prediction.

cross entropy loss given a joint prediction in classification) but may not be the most task-relevant, which means that the weaker modality is not pushed to learn more as the gradient norm for the joint embedding is near zero. This biases the model to mostly depend on this dominant modality during inference. Our method takes an ensemble approach, leveraging a unimodal loss per unimodal backbone to encourage each model to learn its respective unique features as best as it can without having to depend on other modalities, which leads to better overall multimodal predictions when fused (e.g. weighted average of unimodal logits) during inference time.

III. Overview

A. Problem Formulation

First, we restricted the definition of multimodal data in this study to mean textual and image data which are presented together to form information which is consumed by a person. We would like to define the concept of fake news in this study, utilizing the descriptions of the datasets.

The Fakeddit dataset [13] requires that fake news fall under one of five categories: Satire, Misleading Content, Imposter Content, False Connection, or Manipulated Content.

- Satire: True content is phrased in a satirical way that makes it false.
- Misleading Content: Manipulated information with intent to trick people.
- Imposter Content: Bot-generated content.
- False Connection: Images do not connect to the text which it is paired with.
- Manipulated Content: Images are photoshopped.

The Weibo dataset [7] classified fake news based on tweets which were reported and investigated by a committee.

We want to detect fake news, turning this into a binary classification task where we utilize machine learning to classify a sample as either fake news or real news. This is the problem we try to solve in this study.

B. Solution Overview

Our proposed solution utilizes a Unimodal Ensemble (UME) architecture to incorporate both textual and visual information. The solution architecture consists of separate image and text processing pipelines. Each input modality is passed through its corresponding backbone and classifier to obtain unimodal output logits. These logits represent the individual predictions of the image and text models regarding the verifiability of the news content, which are then evaluated with their respective binary cross entropy loss. We then directly average the logits to combine the unimodal predictions, creating joint logits used in the final inference.

A key aspect of our approach is the separate but parallel training of each unimodal model rather than a joint training strategy. By independently training image and text models, we allow the gradient descent optimization to bring each unimodal model to its maximum potential in learning to predict fake or real news. Each modality focuses on its task-relevant features and patterns without being influenced by the other modality during the learning process.

Compared to fusion-based models that attempt to learn cross-modal interactions and correlations, our Unimodal Ensemble approach offers several advantages. First, it mitigates the problem of modality laziness, as each modality is equally important and trained to make predictions independently. This intentional design decision ensures that both textual and visual information are both effectively used in the fake news detection task. Second, by avoiding the need for complex fusion mechanisms and deviations from the standard training pipeline, our solution remains computationally efficient and straightforward, making it easy to adopt to various datasets.

IV. Methods

To introduce formal notation, in this supervised learning problem, we are given a multimodal dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i = (x_{i1}, x_{i2})$ represents the input data with two modalities, x_{i1} and x_{i2}, and y_i is the corresponding output or target variable. We have two modalities, x_1 and x_2, for text and image data, respectively, and unimodal models for each will extract relevant features for downstream fusion and inference. We will denote $f(x_1)$ as the unimodal text encoder and $g(x_2)$ as the unimodal image encoder. For simplicity, we will assume that a linear classifier is included in each unimodal encoder function to output binary prediction logits.

A. Architecture Choices

Since we were computationally limited in scope, we chose to use transfer learning by initializing our unimodal backbones for each of image and text using strongly pretrained weights, particularly those acquired from Contrastive Language-Image Pretraining (CLIP) [18].

We employ CLIP-variants as the image and text unimodal backbones in our ensemble, specifically using SigLIP [26] for the Fakeddit dataset and Chinese CLIP [25] for the Weibo dataset. This choice was made to capture the semantic and contextual information from the news articles and associated images in a domain-specific manner with proper language support. For instance, the text data in the Weibo dataset (see Section V.A for more information) consisted of chinese characters, which is not in the typical training corpora of English-based CLIP models. This called for a special chinese tokenizer and pre-trained model that handled this domain, which was the justification in choosing Chinese CLIP.

B. Training Pipeline

Once the pre-trained weights were loaded, the models were then fully unfrozen and fine-tuned on each of fake news datasets. The unimodal forward pass can be represented as:

$$z_{i1} = f(x_{i1}) \tag{1}$$

$$z_{i2} = g(x_{i2}) \tag{2}$$

$$z_i = \frac{z_{i1} + z_{i2}}{2} \tag{3}$$

$$p_i = \sigma(z_i) \tag{4}$$

where $z_i \in \mathbb{R}^{nx2}$ is the joint logit after directly averaging each unimodal logit. These joint logits are then passed through a softmax function $\sigma(\cdot)$ to attain the joint prediction p_i. Although we use this joint prediction during inference time, we do not directly use it to generate gradients for gradient descent, as this would cause the modality laziness problem in the late-fusion setting [29]. To enforce each modality to completely learn its respective predictive features, we take a unimodal ensemble approach by which each unimodal encoder would have its own loss function. Let L be the standard (binary) cross entropy loss for classification. We can attain unimodal predictions from the unimodal logits shown in equations 1 and 2, which would then be separately evaluated:

$$L_1 = \sum_{i=1}^{n} L(\sigma(z_{i1}), y_i) \tag{5}$$

$$L_2 = \sum_{i=1}^{n} L(\sigma(z_{i2}), y_i) \tag{6}$$

where n is the batch size. From multi-task learning, the total loss function would be additive combination of the unimodal losses:

$$L_{total} = L_1 + L_2 \tag{7}$$

C. Training Details

All experiments were run with Ubuntu 22.04 on a machine with an NVIDIA RTX 4090 GPU with 24GB VRAM. For modular code and efficient training, we leveraged the PyTorch Lightning [6] training framework with fp16 mixed precision and 12 CPU workers. To monitor training and assess evaluation, we used Weights and Biases for logging. For all runs, we used the SGD optimizer with a momentum of 0.9 and weight decay of 1.0e-4.

For the Fakeddit dataset, we used a batch size of 144 samples with a learning rate of 0.01. This learning rate was modulated by a StepLR scheduler that halved the learning rate (i.e. gamma=0.5) every 500 steps. Only one epoch was used for training (similar to what is done in large language model training) since there were sufficiently many training samples necessary for convergence.

For the Weibo dataset, we used a batch size of 50 samples with a learning rate of 0.01. We also used a StepLR scheduler here that halved the learning rate every 250 steps. The model was then trained for 20 epochs with validation metrics logged between every epoch.

Since there is some class imbalance as seen in Table I, we accounted for this during training by balancing out the sampler using the provided PyTorch WeightedRandomSampler class. The robust metrics utilized during evaluation in Section V.C also accounted for this fact.

V. Evaluation

A. Data

We use the Fakeddit dataset [13] and Weibo dataset [7] to evaluate our model. Both are public datasets which contain labeled samples of textual data (Fakeddit is English, Weibo is Chinese) paired with an image. Such a sample is given a binary classification of Fake News (label 0) or Real News (label 1). Both datasets were pre-processed to centralize all the metadata and text data into a single comma-separated values (CSV) file, of which each row contained a sample's text data, supervised label, and a corresponding ID for the respective image file in the image folder directory. Because the CLIP model was pretrained with a maximum of 77 text token length, we first tokenized the text accordingly by either padding or truncating to that length to effectively leverage parallelism in training.

The multimodal subsection of the Fakeddit dataset is a very large dataset (around 110GB data) consisting of around 680,000 samples of text and images scraped from various subreddits off the Reddit social media platform.

This data is split into training, validation, and testing splits. The training split is around 82% of the dataset and consists of 61% Fake News and 39% Real News. This adds up to more than 560,000 samples. The validation and testing sets are approximately the same size and follow the same distribution as the training split. Each are around 60,000 samples in total.

The Weibo dataset is much smaller and is scraped from the Chinese social media platform, Weibo. The labels are given by the official ”rumor debunking system” of the platform. This multimodal dataset consists of more than 9,000 samples with Chinese characters paired with images.

This data is only split into two splits, so we use the smaller split as both validation and testing. The training split consists of around 80% of the dataset (7,481 samples) with fake news making up around 80% of samples. The other split (Validation and Testing) consists of 1,930 samples with fake news making up around 77% of samples.

B. Baselines

We compare our solution against the following models which have also benchmarked their algorithms on the same multimodal fake news detection task. EANN [24], SpotFake

TABLE I
DATASET STATISTICS

Dataset	Split	Total Samples	Fake News	Real News
Weibo	Training	7,481	6,044 (˜80%)	1,437(˜20%)
	Validation	1,930	1,480 (˜77%)	450 (˜23%)
	Testing	1,930	1,480 (˜77%)	450 (˜23%)
Fakeddit	Training	563,613	341,532 (˜61%)	222,081 (˜39%)
	Validation	59,299	35,979 (˜61%)	23,320 (˜39%)
	Testing	59,299	35,781 (˜61%)	23,518 (˜39%)

[20], HMCAN [17], and CAFE [5]. These represent the state-of-the-art models in fake news detection and utilize various approaches as described in Section II. All of these models utilize various multimodal approaches to detect fake news, and are the most heavily cited for this area of work.

Our comparative analysis validates our approach and gives relevant comparisons for readers to see our relative performance against the top models out there.

C. Metrics

We evaluate our model on the typical metrics for a classification task: accuracy, precision, recall, and F1 score.

Accuracy measures the overall correctness of the model across all predictions, which is the most natural measurement of performance in this classification task. Precision measures the proportion of true positives within positive predictions given statistics from the confusion matrix. This measures the ability of a solution to identify samples as positive. Recall measures the proportion of predicted positives within actual positive samples, measuring the ability of a solution to catch as many true positives as possible. Finally, the F1 score is a combination of precision and recall which is calculated by:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

.

Precision, recall, and F1 score have separate scores for Fake News and Real News since we are measuring the performance of the model in identifying these and both must perform well for a comprehensive evaluation.

VI. RESULTS

A. Comparison Against State-of-the-Art

We compare our model against the four state-of-the-art models mentioned in previous sections. Table II summarizes the performance of the models on the two datasets.

Our model, unimodal ensemble (UME), has the highest accuracy for both datasets which is a positive indication that our model outperforms the other models. We can see that the margin is slimmer on Fakeddit dataset, but the improvement in accuracy is more than 3% on the Weibo dataset.

As we dive deeper into our evaluation metrics, we see that we outperform all benchmarks on precision and recall for fake news on the Fakeddit dataset and are competitive for the F1 score for fake news. We see a small increase of around 1% for precision and a larger increase in recall (more than 2%). For real news, we score very closely to the other models, but we do not outperform the state-of-the-art models.

For the Weibo dataset, we outperform Fake News metrics by a significant margin. Precision is increased by more than 3% from the best model, and F1 score is 2% higher than the best model. We also quite competitive with the results of SpotFake in the recall metric. For the real news performance, we do not outperform the other models, but still have comparable performance.

B. Ablation Study

As a part of this study, we experiment with various fusion techniques that attempt to counteract the problem of modality laziness that our solution primarily addresses.

We first baseline with using late fusion, which we know suffers from modality laziness and serves as a lower bound on multimodal performance. We then take a closer look by benchmarking with two proposed solutions: on-the-fly gradient modulation with generalization enhancement (OGM-GE) [15] and quality-aware multimodal fusion (QMF) [28]. As a brief summary, OGM-GE is able to identify the dominant modality using a gradient-based method and artificially slows down the learning rate of the dominant modality so it doesn't immediately saturate the loss and the weaker modality has a chance to learn. QMF does almost the opposite by identifying ”hard” samples through loss trajectories and encouraging the multimodal model to learn more from those samples to prevent loss saturation. Finally, we show the results of our method of unimodal ensembling (UME). We compare the results of these methods in Table III.

As you can see, UME performed the best across almost all metrics across both datasets. The next best was QMF with the best fake news precision and real news recall on the Fakeddit dataset; however, it is important to note that the QMF training collapsed on the Weibo dataset, so we lack true results in that portion of our results. We hypothesize that this is due to the smaller size of the dataset leading to representational collapse.

VII. DISCUSSION

A. Limitations

One limitation of our Unimodal Ensemble architecture is the lack of synergistic interactions between modalities since there it no further depth-wise processing beyond the point of fusion. While this does not seem to hold back performance based on our benchmarking efforts, there is definite room for improvement to better model the case of modality disagreement (i.e. synergy) in the real world, which commonly seen in the task of fake news detection.

This leads us to another limitation of our work. Since we only evaluated the model on two datasets, the robustness of the model in the real world (integration into cyber-physical systems) is unknown. The model may simply learn the features of the datasets very well during training (data dependent), resulting in the excellent results that we present in this paper. Overall, the term ”fake news” is quite subjective, so the task

TABLE II
PERFORMANCE OF DIFFERENT MODELS ON TWO DATASETS.

Model	Fakeddit							Weibo						
	Acc	Fake News			Real News			Acc	Fake News			Real News		
		P	R	F1	P	R	F1		P	R	F1	P	R	F1
EANN	0.724	0.727	0.719	0.723	0.722	0.729	0.726	0.782	0.827	0.697	0.756	0.752	0.863	0.804
SpotFake	0.819	0.801	0.848	0.824	0.839	0.790	0.813	0.892	0.902	**0.964**	0.932	0.847	0.656	0.739
HMCAN	0.881	0.880	0.882	0.881	**0.882**	0.880	0.881	0.885	0.920	0.845	0.881	**0.856**	**0.926**	**0.890**
CAFE	0.912	0.946	0.886	**0.959**	0.878	**0.942**	**0.909**	0.840	0.855	0.830	0.842	0.825	0.851	0.837
UME	**0.919**	**0.953**	**0.910**	0.931	0.872	0.932	0.901	**0.927**	**0.957**	0.948	**0.952**	0.833	0.859	0.846

TABLE III
ABLATION STUDY RESULTS.

Method	Fakeddit							Weibo						
	Acc	Fake News			Real News			Acc	Fake News			Real News		
		P	R	F1	P	R	F1		P	R	F1	P	R	F1
Late Fusion	0.891	0.927	0.889	0.908	0.841	0.893	0.867	0.921	0.950	0.947	0.948	0.826	0.836	0.826
OGM-GE	0.865	0.885	0.891	0.888	0.832	0.825	0.828	0.915	0.950	0.939	0.944	0.805	0.836	0.820
QMF	0.917	**0.955**	0.904	0.929	0.865	**0.935**	0.899	0.767	0.767	-	0.868	0	0	0
UME	**0.919**	0.953	**0.910**	**0.931**	**0.872**	0.932	**0.901**	**0.927**	**0.957**	**0.948**	**0.952**	**0.833**	**0.859**	**0.846**

itself is not well-defined. However, we make a best effort to capture nuances by using multi-lingual datasets from different cultural perspectives in this work.

B. Implications

Since our model outperforms state-of-the-art models, we can infer that there is a gap in the understanding of multimodal model training which necessitates correction. If separating the modalities can increase performance, it means that there is some unimodal information lost in the fusion-based training approaches, and the field is not fully leveraging all modalities to the degree that they can be learned.

On the contrary, this finding also implies that there are task-relevant, intra-modal signals for fake news detection, meaning that within each modality, there are patterns which discern fake news from real news. These signals are present within each modality separately, and allows our unimodal ensemble to accurately detect fake news without any cross-modal interactions. Following this line of logic, fusing these unimodal signals (e.g. averaging) allows the multimodal model to be more robust due to the effect of ensembling.

Another interesting finding is that our model performs better on detecting fake news than detecting real news, implying that media inherently might be a biased data source, which could impact the understanding of fake news detection.

While our work provides promising results in detecting fake news, we also want to point out that any deployment of this system in a real world context would require much more testing and human-in-the-loop capabilities to ensure that First Amendment rights are not infringed upon and that bias is minimized.

C. Future Work

In the future, our technique can be applied to other modalities to see if it can consistently outperform other models (or at least achieve similar performance). In the realm of fake news detection, modalities of interest include video and audio, since those two have become deeply integrated into modern social media and the internet, in general. We did not include video as a modality in this initial study, as we believe that most media intake in an academic setting comes from text and images. This would have added additional complexity and computational burden that we leave for future work.

We can also apply the model architecture to other multi-modal problems such as the sign language translation problem [27] and smart classrooms [4]. Since sensor data is even less studied than textual and image data, we could see interesting results in overcoming modality laziness.

Lastly, we must evaluate our approach on more datasets to ensure robustness of the unimodal ensemble approach. This will ensure that the performance translates across many settings and is not specific to these two datasets.

VIII. CONCLUSION

In this work, we have proposed a novel Unimodal Ensemble (UME) architecture for multimodal fake news detection, aiming to tackle the issue of modality laziness in existing fusion-based approaches. By training the unimodal image and text models in independently and in parallel, each modality learns task-relevant features without being influenced by the other during the learning process. This approach ensures that both textual and visual information are effectively utilized in the fake news detection task.

We evaluated our UME model on two widely used datasets, Fakeddit and Weibo, and compared its performance against state-of-the-art models like EANN, SpotFake, HMCAN, and CAFE. The results demonstrate that our model outperforms these baselines regarding overall accuracy across both datasets. Furthermore, our model performs strongly in detecting fake news, as evidenced by its high precision, recall, and F1 scores.

Our ablation study, benchmarking existing solutions for the modality laziness problem, further validates the effectiveness of the UME approach. UME consistently outperforms late fusion, OGM-GE, and QMF across most metrics on both datasets, highlighting the benefits of ensembling separately-trained unimodal models.

As the field of multimodal learning continues to evolve, we believe that our work contributes to the ongoing effort to combat the spread of fake news and misinformation in the digital age.

Acknowledgment

Thank you to Dr. Meiyi Ma for her mentorship and guidance throughout this semester. This project would not have been possible without her.

References

[1] Tankut Akgul, Tugce Erkilic Civelek, Deniz Ugur, and Ali C. Begen. Cosmos on steroids: a cheap detector for cheapfakes. In *Proceedings of the 12th ACM Multimedia Systems Conference*, MMSys '21, page 327–331, New York, NY, USA, 2021. Association for Computing Machinery.

[2] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

[3] Shivangi Aneja, Chris Bregler, and Matthias Nießner. Cosmos: Catching out-of-context misinformation with self-supervised learning, 2021.

[4] Wilson Chango, Juan A Lara, Rebeca Cerezo, and Cristóbal Romero. A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(4):e1458, 2022.

[5] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022*, pages 2897–2905, 2022.

[6] William Falcon and The PyTorch Lightning team. PyTorch Lightning, March 2019.

[7] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816, 2017.

[8] Srijan Kumar and Neil Shah. False information on web and social media: A survey, 2018.

[9] Tuan-Vinh La, Minh-Son Dao, Duy-Dong Le, Kim-Phung Thai, Quoc-Hung Nguyen, and Thuy-Kieu Phan-Thi. Leverage boosting and transformer on text-image matching for cheap fakes detection. *Algorithms*, 15(11), 2022.

[10] Jun Li, Yi Bin, Jie Zou, Jiwei Wei, Guoqing Wang, and Yang Yang. Cross-modal consistency learning with fine-grained fusion network for multimodal fake news detection. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, MMAsia '23, New York, NY, USA, 2024. Association for Computing Machinery.

[11] Shangdong Liu, Xiaofan Yue, Fei Wu, Jing Sun, Yujian Feng, and Yimu Ji. Semantic distillation and structural alignment network for fake news detection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6620–6624, 2024.

[12] Michael Luca and Georgios Zervas. Fake it till you make it: Reputation, competition, and yelp review fraud. *Manage. Sci.*, 62(12):3412–3427, dec 2016.

[13] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*, 2019.

[14] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, MAD '23, page 36–44, New York, NY, USA, 2023. Association for Computing Machinery.

[15] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022.

[16] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 153–162, New York, NY, USA, 2021. Association for Computing Machinery.

[17] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162, 2021.

[18] 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. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[19] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10):13915–13916, Apr. 2020.

[20] 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)*, pages 39–47. IEEE, 2019.

[21] Borhan Uddin, Nahid Reza, Md Saiful Islam, Hasib Ahsan, and Mohammad Ruhul Amin. Fighting against fake news during pandemic era: Does providing related news help student internet users to detect covid-19 misinformation? In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, page 178–186, New York, NY, USA, 2021. Association for Computing Machinery.

[22] Xiangyu Wang, Min Zhang, Weiguo Fan, and Kang Zhao. Understanding the spread of covid-19 misinformation on social media: The effects of topics and a political leader's nudge. *Journal of the Association for Information Science and Technology*, 73(5):726–737, 2022.

[23] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 849–857, New York, NY, USA, 2018. Association for Computing Machinery.

[24] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857, 2018.

[25] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022.

[26] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[27] Qian Zhang, JiaZhen Jing, Dong Wang, and Run Zhao. Wearsign: Pushing the limit of sign language translation using inertial and emg wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–27, 2022.

[28] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR, 2023.

[29] Yedi Zhang, Peter E. Latham, and Andrew Saxe. A theory of unimodal bias in multimodal learning, 2023.

[30] Peng Zheng, Hao Chen, Shu Hu, Bin Zhu, Jinrong Hu, Ching-Sheng Lin, Xi Wu, Siwei Lyu, Guo Huang, and Xin Wang. Few-shot learning for misinformation detection based on contrastive models. *Electronics*, 13(4), 2024.