Inter-modality Discordance for Multimodal Fake News Detection

Shivangi Singhal, Mudit Dhawan, Rajiv R Shah
{shivangis,mudit18159,rajivratn}@iiitd.ac.in
IIIT Delhi
India

Ponnurangam Kumaraguru
pk.guru@iiit.ac.in
IIIT Hyderabad
India

ABSTRACT

The paradigm shift in the consumption of news via online platforms has cultivated the growth of digital journalism. Contrary to traditional media, lowering entry barriers and enabling everyone to be part of content creation have disabled the concept of centralized gatekeeping in digital journalism. This in turn has triggered the production of fake news. Current studies have made a significant effort towards multimodal fake news detection with less emphasis on exploring the discordance between the different multimedia present in a news article. We hypothesize that fabrication of either modality will lead to dissonance between the modalities, and resulting in misrepresented, misinterpreted and misleading news. In this paper, we inspect the authenticity of news coming from online media outlets by exploiting relationship (discordance) between the textual and multiple visual cues. We develop an inter-modality discordance based fake news detection framework to achieve the goal. The modal-specific discriminative features are learned, employing the *cross-entropy loss* and a modified version of *contrastive loss* that explores the inter-modality discordance. To the best of our knowledge, this is the first work that leverages information from different components of the news article (*i.e.*, headline, body, and multiple images) for multimodal fake news detection. We conduct extensive experiments on the real-world datasets to show that our approach outperforms the state-of-the-art by an average *F1-score* of *6.3%*.

CCS CONCEPTS

• **Applied computing** → **Investigation techniques**.

KEYWORDS

Multimodal Fake News, Inter-modality Discordance, Contrastive Loss, Multitask learning, Metric learning

ACM Reference Format:
Shivangi Singhal, Mudit Dhawan, Rajiv R Shah and Ponnurangam Kumaraguru. 2021. Inter-modality Discordance for Multimodal Fake News Detection. In *ACM Multimedia Asia (MMAsia '21), December 1–3, 2021, Gold Coast, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3469877.3490614

MMAsia '21, December 1–3, 2021, Gold Coast, Australia

ACM ISBN 978-1-4503-8607-4/21/12…$15.00
https://doi.org/10.1145/3469877.3490614

1 INTRODUCTION

Fake news is not a new notion. Before the era of digital technology, the circulation of fake news transpired mainly through yellow journalism. It resurfaced during the 2016 U.S. presidential election and till date we have witnessed numerous instances where it transcend into our lives and left with a mark. The Internet and social media changed the ways by which fake news is fabricated and propagated [1]. The paradigm shift in the news consumption via social networks is noteworthy and innumerable number of efforts have been made to curb fake news online [6, 13, 27, 28, 30, 32]. However, little is explored about the news consumption via untrustworthy websites and its consequences in the real world. A recent study by [10] shows that *44.3%* of Americans visited fake websites during the election period. It also showed various evidences that reveal the engagement of users with the articles they came across and were vulnerable to believing the information contained in such claims.

Typically, a news article consists of a headline, content, top-image and other corresponding images. Critical examination of news credibility in presence of such multiple cues becomes challenging. The reasons are two fold. First, the narrative of the news is documented with supporting claims and evidences that makes it lengthy. This in turn open gates for adding fake facts without being getting noticed. Second, news on online media websites are aided with multiple visuals to make it look agreeable. This gives a plethora of opportunities to manipulators to sell their bogus narratives by proving supportive images. Previous research made an attempt to detect fake news by leveraging information from both the modalities [6, 13, 27, 28, 30, 32]. However, limited attention is drawn towards other multiple visual signals that are present in a news sample. Specifically, such methods focused only on the *first-image* in tandem with the textual cue to perform multimodal fake news detection. In addition, the initial learning obtained for each modality is combined in an additive manner, ignoring the relationship across modalities for fake news classification.

In this paper, we focus on fixing the above mentioned drawbacks by exploiting information from all the graphical cues aggregated with the textual details. We believe incorporating multiple images is beneficial due to the following reasons: *(i)* understanding story in a text often requires reader to develop mental imagery skills [16, 33]. Images can facilitate creation of such mental representations [8] and can result in deeper learning [17, 23, 24], *(ii)* images assist in clarification of ambiguous relations in the text, often termed as "multimedia effect" [17], *(iii)* while words can be viewed as descriptive representations, images, in contrast, depict the external representations, showcasing the meaning that the text represent [2].

Our paper aims to capture the synergies between the modalities for multi-modal fake news detection based on inter modality discordance score. We hypothesize that fabrication introduced in

any modality will lead to dissonance between them *i.e.* the obtained feature vectors from a fake (real) sample, when projected in a multi-modal space, will be distant (closer) and portrays negligible (significant) relationship between the involved modalities [4]. We examine the discordance score based on a modified version of the *contrastive loss* that enforces distinct features of a real sample to be closer to each other and farther for fake news. The designed method is also able to classify samples comprising of only unimodal features as the modality specific sub-modules are able to independently learn discriminative features via the imposition of the *cross-entropy loss*. The main contribution of our work is summarized below.

- We present a novel framework that leverages information from multiple images in tandem with the text modality to perform multimodal fake news detection. The count of images varies on per sample basis and our designed method is able to incorporate such changes efficiently.
- We adopt a multimodal discordance rationale for multimodal fake news detection. Our proposed model effectively captures both the intra and inter modality relationship between the different modalities.

2 RELATED WORK

Existing studies have investigated textual information present in a news article for fake news detection by analysing either the linguistic styles or lexical features [19]. Some have also experimented with the sentences present in the text with BiLSTM [12] and vector space model [22]. In addition to text, a news article also comprises of visual information that can be leveraged for fake news detection. In recent times, we have witnessed a substantial growth towards content based multimodal fake news detection that combines information from both the text and the corresponding image. In this section, we revisit the works that placed emphasis on incorporating either single or multiple images in tandem with the text for fake news detection.

2.1 Multimodal Single Image Fake News Detection

With recent advancement in the technology, manipulators have become more experimental in creating deceptive stories. To counter proliferation of such manipulated content, studies have exploited both the textual and visual information. Depending on the length of an article, existing research can further be grouped into two categories. One focusing on social media *i.e.* the short-length datasets and other pertaining to news from online news websites *i.e.* long-length datasets. For short-length datasets, Wang *et al.* [30] designed an end-to-end framework that aims to detect fake news by discarding the event specific features and emphasising on the available shared features, in addition to the textual and visual signals. Khattar *et al.* [13] established correlation across the modalities by designing a multimodal variational autoencoder. The module performs reconstruction of representations from both the modalities via a learned shared feature vector. The method is used in tandem with the classification module to detect fake news. Singhal *et al.* [28] leveraged contextual text information combined with the image features to perform multimodal fake news classification. Here, the primary goal is to remove the secondary sub-tasks *i.e.* event-discriminator [30] and capturing correlations [13] from the previous discussed approaches to perform multimodal fake news detection. For long-length articles, an advanced version of [28] was proposed by Singhal *et al.* [27]. The method performed an additive fusion of the textual and the visual features extracted fraom a pre-trained XLNet [31] and VGG-19 [26] respectively. Recently, Cui *et al.* [6] focused on exploiting the user comments obtained on a particular post. They designed an adversarial network to capture and preserve the sentiments present in the comments to distinguish fake news effectively. Lately, similarity aware fake news detection was presented by Zhou *et al.* [32]. The paper investigates the relationship between multiple modalities present in a news sample to classify it as fake or real. To capture the relationship effectively, images are first converted into text using pre-trained image2sentence model [29]. Next, the relationship between the modalities is captured by performing a modified version of cosine similarity.

2.2 Multimodal Multiple Image Fake News Detection

Recently, a single study by Giachanou *et al.* [9] introduced a new direction by exploiting the information from multiple images in accordance with the headline and the complementary *first image*. Giachanou *et al.* [9] proposed a multimodal multi-image module that encapsulates information from multiple images in the form of tags and semantic features via a pre-trained VGG-16 network. Next, to establish similarity between the different components of the two modalities, cosine similarity score is calculated between the text and image tags. Finally, textual and visual feature vectors are combined with the similarity score, in an additive manner to perform fake news detection.

2.3 Analysis of the Related Work

Upon examining the related literature, we find the strongest baselines for single-image and multi-image content-based multimodal fake news detection to be SAFE [32] and Giachanou *et al.* [9] respectively. Next, we present remarks on the strongest baselines to situate our work with respect to them.

- In the research presented by Zhou *et al.* [32], *(i)* the textual features are extracted via a Text-CNN [15] ignoring the contextual information, *(ii)* the image is converted into text via image2sent [29] model. Next, cosine similarity is calculated to explore the relationship between the two modalities. We believe converting an image into text might result into loss of semantic information within an image and, *(iii)* no comparison is shown with the existing state-of-the-art methods to demonstrate the effectiveness of the proposed model.
- On the other side, work performed by Giachanou *et al.* [9], lack the reasoning for utilizing multiple images for multimodal fake news classification. Second, taking cues only from the headlines, ignoring the content might lead to information loss. Third, while capturing the similarity, top ten image tags are preferred over the image features. This might lead to inconsistent results as, *(i)* extracted tags might fail to capture the semantic relationship across the images, *(ii)* incorporating only top ten tags might not capture the information present in the image effectively and, *(iii)* extracted

tags might be limited by the vocabulary of the pre-trained model used for extraction and can introduce external bias in the final representations.

To address the above mentioned issues, we present an inter-modality discordance based multimodal fake news detection method. It captures intra-modality relationship by extracting the sequential information from both text and multiple images. In addition, it also forms a multimodal representation of the news article to explore the hidden latent patterns. Our work also introduces a novel application of contrastive loss, employed for measuring the discordance between the components. Enforcing all such losses in conjunction enables for a better feature extraction and robust learning to achieve state-of-the-art performance on the multi-image multimodal fake news detection. Next, we discuss in detail the components of our proposed model.

3 PROBLEM FORMULATION

Assume we have a set of n news articles, $N=\{(H_i, C_i, V_i, y_i)\}_{i=1}^{n}$. Each news sample N_i consists of four elements, *i.e.* headline (H_i), text content (C_i), image-set (V_i) and the corresponding ground-truth label y_i. We formulate the problem as a binary classification task where N_i can be categorized as either fake *(y=1)* or real *(y=0)*. Specifically, our aim is to combine complementary information from multiple modalities for fake news detection in online news websites. Existing methods opted for diverse range of solutions to detect fake news. For example, *(i)* works [27, 28] extracted discriminative features from each modality and performed multimodal fusion to obtain the resultant news vector, *(ii)* other works [13, 30] added a complementary task to perform fake news detection, *(iii)* recent work [9, 32] attempts to exploit the relationship between text and image modalities for fake news detection. All these works show the benefits of leveraging unimodal features, adding complementary tasks and studying relationship for multimodal fake news detection.

Taking cues from all the above mentioned approaches, we formulate the problem as a binary multi-task learning method where our primary objective is to perform multimodal fake news detection (section 4.4). The other related atomic (auxiliary) learning tasks are as follows: *(i)* Inter-modality discordance score, that ensures components of a real news article are pulled together in an embedding space, while simultaneously pushing apart the components of a fake news article (section 4.1) , *(ii)* Unimodal multiple-visual feature extractor, that excavates hidden patterns within a set of sequential images, to obtain the final discriminative rich embeddings (section 4.2) and, *(iii)* Unimodal text feature extractor, that embodies the intra-modality relationship via granular fragment representation, independently from the headline and the content (section 4.3).

4 METHODOLOGY

We present a high-level diagram of our proposed approach in Figure 1. The model performs multi-task operations with primary goal being multimodal fake news detection. It comprises of four components, *(i)* Inter-modality discordance score, *(ii)* Text feature extractor, *(iii)* Multiple-visual feature extractor and, *(iv)* Multimodal fake news detector. Next, we discuss each component in detail.

4.1 Inter-modality Discordance Score

The first auxiliary task presented in our proposed method is calculating the discordance score. It captures the relationship (discordance) between various components present in a news article for multimodal fake news detection. More specifically, the idea is that the average distance between the different components of a fake news article is greater than the average distance between the different components of a real news article, in a multimodal space. We believe measuring discordance has the following implications. A recent study by Claire Wardle, *First Draft News Research Director*[1] presents a list of seven different types of fabricated content circulated in the online world. Though all these forms of misinformation are created differently, some of them can be captured by measuring discordance between different components of a news article. For example, capturing relationship will help in easy identification of fake stories where, *(i)* headlines and visuals are not supporting the content, *(i)* genuine content is circulated with false contextual information, *(iii)* both the content and image are real but the context in which they appear frames a false story. Taking inspiration from [14], we measure the inter-modality discordance score via a modified version of contrastive loss function. It is a form of metric learning that has shown significant improvement over the conventional cross entropy loss for supervised classification [5]. The objective is to predict relative distance between the inputs.

Algorithm 1: Measuring Inter-modality Discordance Score (Training Phase)

Input: $P = [H_i^{\mathbb{R}}, C_i^{\mathbb{R}}, \ldots, V_i = \{I_1^{\mathbb{R}}, \ldots, I_k^{\mathbb{R}}\}_{k=1}^{l}]_{i=1}^{n}, y \in (0,1), M$
Output: Loss
for *each* P_i i.e. $(H_i, C_i, \{I_1, \ldots, I_k)$ **do**
  $r_{ce} = \frac{1}{|P_i|} \sum P_i$;
  $distance = \frac{1}{|P|} \sum_{i=1}^{|P|} \left\| r_{P_i}, r_{ce} \right\|$;
  if *y==1* **then**
    $Loss = max(0, M - distance)$;
  else
    $Loss = distance$;
  end
end

The detailed outline to calculate the inter-modality score is summarized in Algorithm 1 where (H_i, C_i, V_i) depicts the intermediate feature representations for the headline, content and image-set respectively. r_{ce} denotes the centroid value and *distance* signify the average distance between the components of a news sample from the centroid. The distance metric chosen for measuring similarity is the euclidean distance (*l2-norm*). M indicates margin value. The function of margin is that, when average distance between the different components of a fake news article are distant enough, no efforts are wasted on enlarging that distance. However, when that distance is not greater than M, then loss will represent a positive value, and net parameters will be updated to produce more distant feature vectors. The vice-versa happens for the real news articles.

[1] https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79

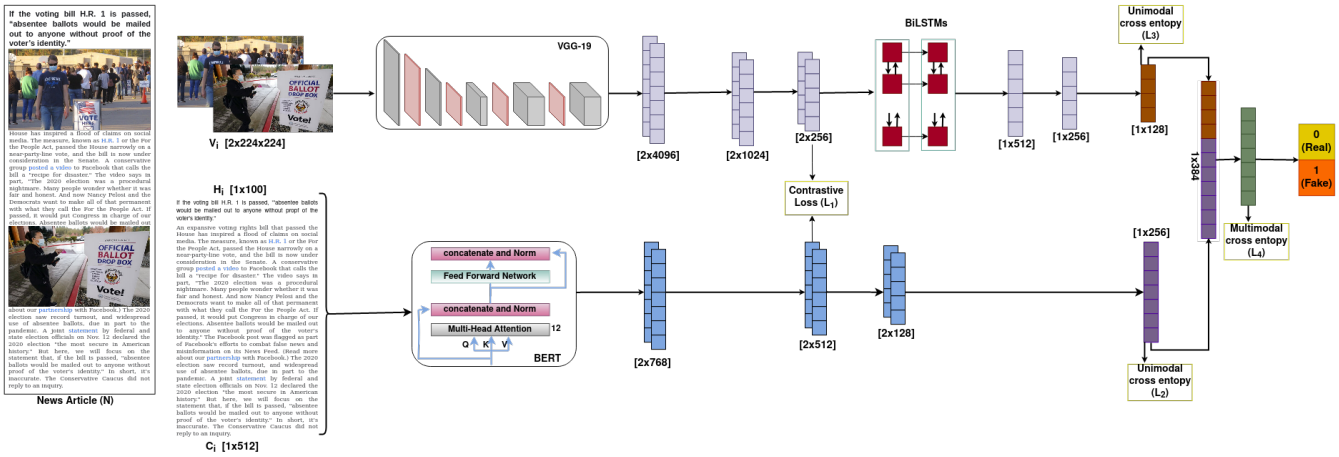

Figure (1) Illustration of the proposed model. It comprises of a primary task *i.e.* multimodal fake news detction. We introduce three auxiliary learning tasks *i.e.* measuring inter-modality discordance score via *contrastive loss*, multiple visual feature extractor and, textual feature extractor.

4.2 Unimodal Visual Feature Extractor

The second auxiliary task considered in the proposed method is the multiple visual feature extractor. Taking inspiration from Giachanou *et al.* [9], we present a novel system that extracts sequential information from multiple images in a two-fold manner. First, the pre-processed images are passed through a VGG-19 network pre-trained on a ImageNet database. The second to last layer of the VGG-19 network serves as a feature embedding for each image present in the news article. Next, to capture the temporal features from the intermediary sequential visual cues, we employ a Bidirectional Long-Short Term Memory (BiLSTM) cells. The continued representations then obtained are passed through fully connected layers to match the length of vector dimensions with that of the resultant textual feature vector.

4.3 Unimodal Text Feature Extractor

The third auxiliary task introduced in the proposed method is the textual feature extractor. It extracts contextual representations from the headline and the content of a news sample. Context refers to information that helps the message of a literary text interpret accurately. Unlike Word2Vec [18] and GloVe [21] which are context insensitive, the word embeddings generated by Transformer [7] are context sensitive representations.

In our work, each content piece (C_i) of a news article is segregated into sentences. A sentence s is represented as a sequence of WordPiece tokens $\{w_s^1, w_s^2, \ldots, w_s^k\}$, where w_s^k is aggregation of the token, position and segment representation for the k^{th} token present in a sentence s. Similarly, headline (H_i) is divided into tokens for further processing. We make use of BERT (Bidirectional Encoder Representations from Transformers) architecture [7]. It is deeply bi-directional that looks at the words by jointly conditioning on both left and right context in all layers.

4.4 Multimodal Fake News Detector

In this section, we discuss the primary task of our proposed method *i.e.* multimodal fake news detection. It leverages information from the textual and multiple visual entities of a news sample to form a multimodal feature vector. Although, we include auxiliary task that extracts modal-specific features from the news article, the necessity to add multimodal features is two fold, *(i)* capturing information from multiple modalities will help in creating a more robust system as compared to the ones build solely on unimodal features and, *(ii)* multimodal features will be more capable in discovering non-trivial patterns and relationship between data instances.

4.5 Loss Functions

The proposed method comprises of a primary task and three other related auxiliary (atomic) learning tasks. We employ a combination of loss functions from all the four tasks to better perform the desired task. It is to be noted that all four tasks are performed during model training but the primary task is considered when assessing the model's performance.

To calculate inter-modality discordance, the training objective is that euclidean distance between the various components of a news sample, in a multi-modal space, is minimised for the genuine news articles and maximized for the false news samples. Taking inspiration from Khosla *et al.* [5], we use the modified version of contrastive loss, originally presented for training of Siamese networks [3] to calculate the inter-modality discordance score. This ensures the distinction between the positive and negative samples effectively. The loss function is represented in Equation 1.

$$L_1 = \begin{cases} \frac{1}{n}\sum_{i=0}^{m} d(r_m, r_{ce}), & \text{if real sample} \\ max(0, M - \frac{1}{n}\sum_{i=0}^{m} d(r_m, r_{ce})), & \text{otherwise} \end{cases} \quad (1)$$

Here, r_m depicts the embedding vector for the *m-th* component of a news article, r_{ce} denotes the centroid, M depicts the margin, set as a hyper-parameter and, $d(\cdot)$ denotes the euclidean distance

between the component of a news sample and its corresponding centroid value.

To capture discriminative unimodal features, we employ the cross-entropy loss to learn the modal independent representations in a robust fashion. In addition, to model the cross-correlations between the entities, we perform a multimodal fusion on intermediate features to form the desired multimodal news vector. The corresponding loss functions for the same is depicted in Fig 1 by *L2*, *L3* and *L4*, respectively.

Hence, the final loss for the proposed method is the weighted sum of the four losses *i.e.*, $L = \alpha L_1 + \beta L_2 + \gamma L_3 + \delta L_4$, where $(\alpha, \beta, \gamma, \delta) \in [0, 1]$. In our experiments, we set the value to be one. However, further hyper-parameter tuning can be performed on these values.

5 EXPERIMENTS AND RESULTS

In this section, we present a series of experiments to demonstrate the efficacy of our proposed method. Specifically, we aim to answer the following research questions:

- **RQ1**: Is the proposed model able to improve multimodal fake news detection by incorporating multiple images?
- **RQ2**: How effective modality discordance hypothesis is in multimodal fake news detection?

Next, we give an overview of the dataset and baseline models, followed by a detailed investigation on the questions being asked.

5.1 Datasets

We use the following publicly available datasets to perform multimodal fake news detection.

- FakeNewsNet Repository (raw version): The dataset introduced by Shu *et al.* [25] comprises of news articles belonging to either political or entertainment discipline. The fake and real news article pertaining to political domain are collected from Politifact[2] whereas fake and real samples for the entertainment domain are gathered from GossipCop[3] and *E*! Online[4] respectively. There are 432 and 624 samples in Politifact for fake and real classes respectively. Likewise, there are 5,323 and 16,817 samples in GossipCop for fake and real classes respectively.
- FakeNewsNet Repository (clean version): Giachanou *et al.* [9] performs multimodal fake news detection by using a portion of dataset released by Shu *et al.* [25] *i.e.* considered GossipCop dataset. Next, they performed dataset cleaning in which all the news samples with non-news content images are removed by performing dedupliacton and manual intervention. In our study, for a fair comparison with the state-of-the-art, we used the cleaned version provided by the authors [9] that consist of *2,745* fake and *2,714* real samples having at least one image associated with them.

[2]https://www.politifact.com/
[3]https://www.gossipcop.com/
[4]https://www.eonline.com/

5.2 Baselines

We compare our proposed methodology with a representative list of state-of-the-art multimodal fake news detection algorithms listed as follows:

- LIWC [20]: It stands for Linguistic Inquiry and Word Count. The method is used to classify the text samples along the psychological dimensions. It identifies how much percentage of the words present in the text lies into any of the linguistic, psychological, and topical categories. Such analysis of the data is then fed as an input the model for further analysis.
- VGG-19 [26]: VGG19 is a variant of VGG model that comprises of 19 layers. Here, we use a fine-tuned version of VGG-19 as a baseline for images.
- Att-RNN [11]: The method is designed to utilize the textual, visual and social-context features for fake news detection. The variant of the model used in the paper excludes the social-context information for a fair comparison.
- SAFE [32]: The objective of this model design is to capture the similarity among modalities to jointly exploit the multimodal information and excavate better representations for multimodal fake news detection. For this, a modified version of cosine similarity is introduced. The text and visual features are extracted by passing the initial representations through Text-CNN [15]. The intermediate representations for the images are obtained via image2sentence model.
- Multi-image Multimodal Method [9]: This is the first research that explores multiple images in tandem with text to perform fake news detection. To extract visual features from multiple images, tags information in combination with the features obtained via pre-trained VGG19 network is used. Authors also exploit semantic information *i.e.* text-image similarity by calculating cosine similarity between them the modalities.
- L2: It is a variant of the proposed model when using only the textual information.
- L3: It is a variant of the proposed model when using only the visual information.
- L2+L3+L4: It is a variant of the proposed model without the inclusion of the similarity score, i.e. inter-modality discordance score.

We compare the performance of our proposed approach with the single-image and multi-image multimodal fake news detection state-of-the-art methods. Currently, SAFE [32] and Giachanou *et al.* [9] serves as the strongest baselines for the single-image and multiple-image respectively. Additionally, we also demonstrate the importance of each component in the proposed method by performing the ablation study.

5.3 Multimodal Fake News Detection RQ1

To answer the question, we compare our proposed method with the existing state-of-the-art models described in Section 5.2. A comparative table depicting the results and improvements of the proposed methodology with the strongest baselines is shown in Table 1 and Table 2 respectively. We draw the following inferences:

- From Table 1, we observe that our proposed model beats the text only and image only baselines for both the datasets.

- From Table 2, we observe that our proposed model beats the strongest baseline for multi-image multimodal fake news detection [9].
- Additionally, for a comparison with the single-image multimodal fake news detection methods, our proposed model outperforms att-RNN [11] and SAFE [32] on the Politifact dataset. This is shown in Table 1. However, we observe inconsistent performance on the GossipCop (raw) dataset.

Investigating the inconsistent performance of the proposed model on Gossipcop (raw) Since previous studies [9, 27] have pointed out the presence of non-news images (*i.e.* logo, gifs, icons of the news websites and advertisements) within the news articles for the GossipCop (raw) dataset. We hypothesize that existence of such noisy images lead to decrease in the performance of our proposed method. To investigate the case, we first performed intersection on the GossipCop (raw) and GossipCop (clean) datasets to get the resultant set comprising of non-news images. We observe that on an average for a sample, 77.77% of the images are non-news in the raw dataset. We further examined what amount of noise is passed through the model. We found that on an average for a sample, there is a *0.8* probability *i.e.* atleast two out of three images passed to the model will be noisy. This probability further shoots up to *0.97* for atleast one out of three images passed to the model. In conclusion, our proposed model is designed to capture the discordance between the modalities. Since there exists no relationship between the incoming noisy images and text, our model fails to learn any representative pattern about the news article leading to inconsistent results.

Table (1) Comparison of our proposed model with the text†, image$^{\mp}$ and single-image multimodal‡ fake news baselines.

		LIWC†	VGG-19$^{\mp}$	Att-RNN‡	SAFE‡	Proposed Method
Politifact (raw)	Acc.	0.822	0.649	0.769	0.874	**0.913**
	F1	0.815	0.720	0.826	0.896	**0.902**
GossipCop (raw)	Acc.	0.836	0.775	0.743	0.838	**0.850**
	F1	0.466	0.862	0.846	**0.895**	0.743

Table (2) Comparison of our proposed model with the multi-image multimodal fake news detection baselines.

		Giachanou *et al.* [9]	Proposed Method
GossipCop (clean)	Acc.		**0.880**
	F1	0.795	**0.915**

5.4 Evaluating Inter-Modality Discordance (RQ2)

In order to answer RQ2, we perform two experiments.

First, we visualize the average distance between the components of a news sample via Kernel Density Estimate (KDE) plots. Figure 2 shows the distribution of discordance score during the training and testing phase for GossipCop (clean) and Politifact respectively. As stated in Algorithm 1, the margin value signifies the radius around the embedding space of a sample. We hypothesize that the average distance between the components of a real (fake) news sample lies closer (farther) to the radius (*M=1*). We observe that on an average,

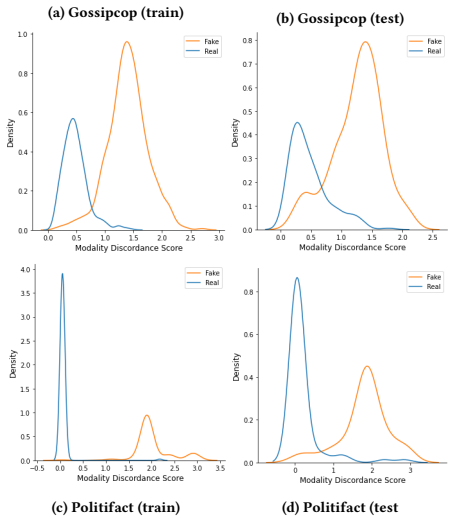

Figure (2) Measuring modality discordance score on train and test set of GossipCop (clean) and Politifact respectively.

the mean distance between the components of real and fake news is *0.476* and *1.39*, respectively, as shown in Figure 2*(a)*. The validity of our results are solidified by the fact that peaks in our density plots are narrow which shows that the variance in the output of a class is low. Additionally, the intersection of area under the curve for real and fake news is minimal indicating a clear separation between the classes. All the aforementioned observation are consistent across datasets and training-validation splits, as depicted in Figure 2*(b-d)*.

Second, we performed an ablation study to examine the performance of our proposed model with its different variants. The results are depicted in Table 3. We observe an improvement in the L2+L3+L4 variant on addition of L1. This shows that adding inter-modality discordance score in the multimodal detection method aids in better fake news detection.

Table (3) Comparison of the proposed model with its different variants.

		L2	L3	L2+L3+L4	Proposed Method (L1+ L2+L3+L4)
Politifact (raw)	Acc.	0.898	0.828	0.906	**0.913**
	F1	0.896	0.821	0.889	**0.902**
GossipCop (clean)	Acc.	0.861	0.684	0.863	**0.880**
	F1	0.906	0.785	0.908	**0.915**

6 CONCLUSION

In this paper, we present an inter-modality discordance based fake news detection method. The method leverages information from both the textual and multiple visuals features of a news sample and, investigates the relationship between them via a modified version of contrastive loss. In addition, cross-entropy loss is enforced on the unimodal and multimodal data streams to ensure that they independently and jointly learn discriminative features. Extensive experiments on the two real-world datasets demonstrate the strong performance of our proposed method.

ACKNOWLEDGMENTS

Shivangi Singhal is supported by TCS Research Scholar Program. Rajiv Ratn Shah is partly supported by the Infosys Center for AI at IIIT Delhi. We also thank Hitkul Jangra, Nidhi Goyal and Ritwik Mishra, members of the Precog Research Lab at IIIT-Hyderabad for the useful discussions.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.

[2] Tobias Bartholomé and Rainer Bromme. 2009. Coherence formation when learning from text and pictures: What kind of support for whom? *Journal of Educational Psychology* 101, 2 (2009), 282.

[3] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*. Springer, 850–865.

[4] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. 2020. Not Made for Each Other- Audio-Visual Dissonance-Based Deepfake Detection and Localization. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 439–447. https://doi.org/10.1145/3394171.3413700

[5] Joon Son Chung and Andrew Zisserman. 2017. Out of Time: Automated Lip Sync in the Wild. 251–263. https://doi.org/10.1007/978-3-319-54427-4_19

[6] Limeng Cui, Suhang Wang, and Dongwon Lee. 2019. SAME: Sentiment-Aware Multi-Modal Embedding for Detecting Fake News. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Vancouver, British Columbia, Canada) *(ASONAM '19)*. Association for Computing Machinery, New York, NY, USA, 41–48. https://doi.org/10.1145/3341161.3342894

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

[8] Alexander Eitel and Katharina Scheiter. 2015. Picture or text first? Explaining sequence effects when learning with pictures and text. *Educational psychology review* 27, 1 (2015), 153–180.

[9] A. Giachanou, G. Zhang, and P. Rosso. 2020. Multimodal Multi-image Fake News Detection. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. https://doi.org/10.1109/DSAA49011.2020.00091

[10] Andrew M Guess, Brendan Nyhan, and Jason Reifler. 2020. Exposure to untrustworthy websites in the 2016 US election. *Nature human behaviour* 4, 5 (2020), 472–480.

[11] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia* (Mountain View, California, USA) *(MM '17)*. Association for Computing Machinery, New York, NY, USA, 795–816. https://doi.org/10.1145/3123266.3123454

[12] Hamid Karimi and Jiliang Tang. 2019. Learning Hierarchical Discourse-level Structure for Fake News Detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3432–3442. https://doi.org/10.18653/v1/N19-1347

[13] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2915–2921. https://doi.org/10.1145/3308558.3313552

[14] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 18661–18673. https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf

[15] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[16] Walter Kintsch. 1991. The Role of Knowledge in Discourse Comprehension: A Construction-Integration Model. In *Text and Text Processing*, G.E. Stelmach and P.A. Vroon (Eds.). Advances in Psychology, Vol. 79. North-Holland, 107–153. https://doi.org/10.1016/S0166-4115(08)61551-4

[17] Richard E Mayer. 2002. Multimedia learning. In *Psychology of learning and motivation*. Vol. 41. Elsevier, 85–139.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 3111–3119.

[19] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.

[20] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[22] Victoria L. Rubin and Tatiana Lukoianova. 2015. Truth and Deception at the Rhetorical Structure Level. *J. Assoc. Inf. Sci. Technol.* 66, 5 (May 2015), 905–917. https://doi.org/10.1002/asi.23216

[23] Wolfgang Schnotz. 2002. Commentary: Towards an integrated view of learning from text and visual displays. *Educational psychology review* 14, 1 (2002), 101–120.

[24] Wolfgang Schnotz and Maria Bannert. 2003. Construction and interference in learning from multiple representation. *Learning and instruction* 13, 2 (2003), 141–156.

[25] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* 8, 3 (2020), 171–188.

[26] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

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

[28] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh. 2019. SpotFake: A Multi-modal Framework for Fake News Detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. 39–47. https://doi.org/10.1109/BigMM.2019.00-44

[29] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and Tell: A Neural Image Caption Generator. *CoRR* abs/1411.4555 (2014). arXiv:1411.4555 http://arxiv.org/abs/1411.4555

[30] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & amp; Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 849–857. https://doi.org/10.1145/3219819.3219903

[31] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237* (2019).

[32] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multi-Modal Fake News Detection. arXiv:2003.04981 [cs.CL]

[33] RA Zwaan and GA Radvansky. 1998. Situation models in language comprehension and memory. *Psychological bulletin* 123, 2 (March 1998), 162—185. https://doi.org/10.1037/0033-2909.123.2.162