MMM: An Emotion and Novelty-aware Approach for Multilingual Multimodal Misinformation Detection

Vipin Gupta[⁑,1], Rina Kumari[⁑,1], Nischal Ashok[2], Tirthankar Ghosal[3], Asif Ekbal[1]
[1]Indian Institute of Technology Patna, India
[2]UMass Amherst, Massachusetts, United States
[3]Institute of Formal and Applied Linguistics, MFF, Charles University, Czech Republic
{vipingupta1907,rinakri08,nischal.ashok09,asif.ekbal}@gmail.com
ghosal@ufal.mff.cuni.cz

Abstract

The growth of multilingual web content in low-resource languages is becoming an emerging challenge to detect misinformation. One particular hindrance to research on this problem is the non-availability of resources and tools. Majority of the earlier works in misinformation detection are based on English content which confines the applicability of the research to a specific language only. Increasing presence of multimedia content on the web has promoted misinformation in which real multimedia content (images, videos) are used in different but related contexts with manipulated texts to mislead the readers. Detecting this category of misleading information is almost impossible without any prior knowledge. Studies say that *emotion-invoking and highly novel* content accelerates the dissemination of false information. To counter this problem, here in this paper, we first introduce a novel multilingual multimodal misinformation dataset that includes *background knowledge* (from authentic sources) of the misleading articles. Second, we propose an effective neural model leveraging *novelty detection* and *emotion recognition* to detect fabricated information. We perform extensive experiments to justify that our proposed model outperforms the state-of-the-art (SOTA) on the concerned task [1].

1 Introduction

Fast adoption of social media platforms have promoted people to knowingly or unknowingly subscribe, create and share misleading, fake, and irrelevant information which consists of various attributes like title, text information, visual information, etc. These attributes may contain false or misleading information. The news or stories having false information is called misinformation. In recent years, we observe substantial advancements in automatic fake news detection. However, most of these are targeted to resource-rich language like English. When it comes to the scenario of (relatively) low-resource Indian languages like Hindi, Bengali and Tamil, the amount of research is insignificant, primarily due to the unavailability of data and other associated resources. With the advancement of multimedia news on the internet, news containing same (non-novel) image with different (novel) text influences the fake news on social media to mislead the newsreaders. Since the image looks authentic and aligns with the new text, it becomes very challenging to detect this category of fake news. The implication of misinformation detection with novelty detection and emotion recognition was first presented by MIT Scholars[2]. Novelty refers to the extent to which news readers encounter unfamiliar news, which may include some element of surprise. In this work, we take forward the misinformation work on the shoulder of novelty detection via entailment task with emphasis to textual similarity measures. Literature also suggest that novel and emotion invoking contents in the news articles act as fuel for the rapid dissemination (Kumari et al., 2021a),(Kumari et al., 2021b) and (Kumari et al., 2022).

Although people have performed an extensive investigation in different dimensions of misinformation detection, however, a very few mechanisms have focused on novelty and emotion aware misinformation detection with background knowledge for the relatively low-resource languages. We make an attempt to address these challenges by creating important resources and effective baseline. We first introduce a novel multilingual multimodal misinformation dataset for the Indian languages like Hindi, Bengali and Tamil. The instances (here, in-

⁑ corresponding author

[1] Code and Data is available here: 1. https://www.iitp.ac.in/~ai-nlp-ml/resources.html#MMM_Dataset, 2. https://github.com/vipingupta1907/MVEN

[2] https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 464–477
November 20–23, 2022. ©2022 Association for Computational Linguistics

stance means a single claim may be real or fake) in each language are different, meaning the same instance will not be present in more than one language. During training the model, we mix the instances of all the three languages which makes it multilingual. We further design a deep learning-based misinformation detection model using novelty detection and emotion recognition as the assisting tasks.

The major contributions offered in this article are as follows:

- We create a novel multilingual multimodal misinformation dataset for Indian languages, which is, to the best of our knowledge, the very first attempt toward creating the corpus for multimodal misleading information detection where the same image is used in a different context to convey false information.
- We propose a multilingual multimodal framework using novelty and emotion recognition as the assisting tasks for misinformation detection, where the main task is to check whether the same image has been published earlier in a different context in other languages.
- We perform zero-shot experiments on our proposed architecture to demonstrate the robustness of our model on the unseen languages at the training time and obtain encouraging performance.

2 Related Work

The concept of misinformation or fake news detection has started in early 2010, as social media started to have an immense impact on people's views. Shu et al. (2017) has introduced one of the first extensive studies for misinformation detection on social media. It has described the fact-checking methodologies as verification of the hypothesis made in a news article to judge if the claim is true or not. The work presented in FakeDetector (Zhang et al., 2020) introduces a deep diffusive network model to detect fake news by learning the representations of news articles, creators, and subjects simultaneously.

Shu et al. (2019) introduced a sentence-comment co-attention sub-network which learns and captures check-worthy sentences and user comments jointly to explain why a particular news piece is detected as fake. People have organized many challenges for fake news detection that introduce several novel mechanisms. A competition *The Fake News Challenge (FNC)*[3] introduced a few works (Slovikovskaya and Attardi, 2020), (Chaudhry et al., 2017) for stance detection which are useful to understand attitudes expressed in texts. Stance detection means the detection of relative perspective of two text fragments. The stance detection justifies whether the news article agrees, disagrees, discusses or is unrelated to the news title. If the news article disagrees or is unrelated to the news title, it indicates a high probability of the news to be fake. Few notable works, such as Yin and Roth (2018) and Nie et al. (2019) verify the human generated claims as fake or real. The mechanism presented by Saikh et al. (2020) depicts a word attention-based deep learning model for automatic fake news detection.

The work explored in Jin et al. (2017) combines the textual, visual, and social context features using an attention mechanism for fake news prediction. In continuation to it, EANN (Wang et al., 2018), VAE (Khattar et al., 2019) and SpotFake Singhal et al. (2019) have introduced deep learning-based models and justified that the model is efficient in handling newly emerged events better than the existing methods. The research explored in (Kumari and Ekbal, 2021; Wu et al., 2021; Song et al., 2021) have given attention to feature fusion along with the feature extraction mechanisms and proved that the model's performance also depends upon the semantic interaction between different modalities. The method explored in Zhang et al. (2021) has taken the first step to find the credibility of previously published news articles on the same events as the background knowledge by introducing the Supervised Contrastive Learning (SCL) (Zhang et al., 2021).

One of the promising works (Abonizio et al., 2020) in multilingual misinformation detection explores language-independent fake news detection, which successfully differentiate fake, satirical, and legitimate news across three different languages. Another multilingual work presented in Guibon et al. (2019) uses the convolutional neural network (CNN) to detect fake news with satire on a multilingual dataset. The works presented in (Li et al., 2020b; Glenski et al., 2019) are the major contributors for multilingual multimodal misinformation detection. They first introduced a dataset which

[3] http://www.fakenewschallenge.org/

includes the instances in languages other than English.

Our work is different from the prior works in the perception that (i). we create a Multilingual Multimodal Misinformation (MMM) dataset with background knowledge for relatively low-resource Indian languages which includes the data instances in Hindi, Bengali and Tamil; and (ii). we design a novelty and emotion aware multimodal multilingual framework for misinformation detection.

3 Data Description and Analysis

Several resources like Twitter (Boididou et al., 2015), Weibo (Jin et al., 2017), TI-CNN (Yang et al., 2018), Fauxtography (Zlatkova et al., 2019), Fakeddit (Nakamura et al., 2020), NewsBag (Jindal et al., 2020), etc. are very eminent to study multimodal misinformation detection problems. People have introduced CoAID (Cui and Lee, 2020), MMCoVaR (Chen et al., 2021) and ReCOVery (Zhou et al., 2020) to tackle the misinformation during COVID-19 infodemic. Aforesaid datasets are only available in English language. Very few datasets such as ArCOV-19 (Haouari et al., 2021) and CHECKED (Yang et al., 2021) are available in the languages other than English. MM-COVID (Li et al., 2020a), MuMiN (Nielsen and McConville, 2022) and FactDRIL (Singhal et al., 2021) are the multilingual multimodal misinformation datasets. However, these datasets do not include background information (where and in which context the news has been published first) of the news articles, which are crucial for misleading misinformation detection. Therefore, we prepare a novel Multilingual Multimodal Misinformation (MMM) dataset which includes 10,473 samples. The developed dataset contains the instances from three different Indian languages *viz.* Hindi, Bengali and Tamil. Each instance of the dataset is in the form of source-target pair. Target is the combination of multimodal Hindi, Bengali and Tamil language instances which claim any information or news. The source is the related background information extracted from different websites corresponding to the target.

3.1 Data Collection

Our prepared dataset contains multimedia news disseminated across the country which are mostly centered around the politics, covid-19, social, health and religion domains. We collect the target instances of our *MMM* dataset in following steps:

Fake Instance Collection: We consider the FactDRIL (Singhal et al., 2021) dataset to prepare fake instance in our dataset. FactDRIL is a multilingual multimodal misinformation repository collected from Indian fact checking websites like *althindi, boomlive, newschecker, etc.*, which includes the instances of claim and their investigations in 13 low-resource Indian languages along with the English language. We only consider the multimodal instances from *Hindi, Bengali* and *Tamil* languages to prepare our dataset. We form a set of target samples by combining these instances which includes the fake claim and image URL pair and assign fake label to all instances.

Real Instance Collection: To collect real data instances, we choose two trusted news websites such as *News18* and *Abplive*. Then we crawl all the pages having general-domain national news and scrape all the news article URLs using request module and beautifulsoup[4] library of python. Using the news article URLs, we again webscrape main news content and image associated with the news articles. We collect only Hindi instances from Abplive website and Hindi, Bengali, Tamil instances from the News18 website. At last, we assign real label to each instance. We collect the background information for each multimodal instance of the target sample set in the following steps:

Source Information Extraction The target instance may have more than one image URL. We use OpenAI CLIP Model (Radford et al., 2021) with Multilingual Knowledge Distillation (MKD) (Reimers and Gurevych, 2020) to find the most relevant image among all the target images. Thus, we keep only single image URL corresponding to each target instance. After that, we perform Google reverse image search using all target image URL to retrieve the source information. We extract all the URLs of sources that contain text or image information related to the target image. Now, we send a get request to all URLs of the sources and then extract the text and images present on that particular source. If there is no source information available, we discard these target instances. In order to make the dataset multimodal, we also discard all the source-target pairs without images. In case of source texts in languages other than the respective target text language, we translate the source text into the target text language using googletrans

[4] https://pypi.org/project/beautifulsoup4/

python library [5]. In order to gather authentic background knowledge, the source itself must be highly credible. So, we evaluate the source credibility in the next step.

Credible Source Selection It is not necessary that the entire news article includes false content, instead that some small portion of the news may have false information. We assume that some websites always publish true news. On the other hand, some websites always publish false news. The trusted news website may also have some misinformation but they are very rare and unintentional. During collection of background information of each instance, we had gone through multiple websites. As per the above discussion, these websites may also contain misinformation. So to consider the source information only from the trusted websites, we have used MediaBias scores of different websites and eliminated the information obtained from non-trusted websites. Here, we use MediaBias score to determine the credibility of the websites from where we collect the data. We don't use this MediaBias score for the credibility checking of the instance. MediaBias assigns a class among the six classes *viz.* very high, high, primarily factual, mixed, low, and very low. We consider maximum four source information only from *very high, high, and primarily factual class*. We limit the number of sources to four because each target instance has, on average, four multimodal source information. We extract textual information from credible websites and save all the images present on these websites. For each instance, we have up to 4 sources where each source has some piece of text and a list of images. We consider the piece of text as the source text. Although the main purpose of this step is to shorten the background information up to four, however some target instances are also removed due to the low credibility of the source. By doing so, we extract textual information from credible source websites and save all the images present on them. Thus we discard all the source information extracted from low credible source websites.

Source Image Selection In this step, we remove all the images having dimension less than 50x50 from the list of images corresponding to each source and subsequently remove the unimodal source information again. We keep only one source image, which is approximately identical to the target image but may have some subtle difference since, our research attempts to detect fake news using non-novel images and novel text. We utilize VGG16 (Simonyan and Zisserman, 2014) and compute cosine similarity to find the similarity between target and source image. As a final step, we preserve only the most similar image from each source.

3.2 Data Annotation

Since we create the *MMM* dataset by collecting real samples from the trusted news sources and fake samples from the existing FactDrill repository, we directly assign the labels as real and fake, respectively. The purpose of the annotation is to keep the source information if it is relevant to the corresponding target instance. Otherwise, it is discarded. Thus, we label every instance with either *yes* or *no*. All instances with *yes* labels are included in the dataset and other instances with *no* label are discarded from the dataset. It is solely based on the textual content of the source and target instances. In addition to automatic annotation, we also perform human annotation to check the quality of automatic annotation.

Automatic annotation We consider two types of annotations for each source-target pair of our MMM dataset:

(i). In the first annotation type, we assign the label of the source-target pair similar to the target label. If the target label is fake, we assign the label as fake and if the target data label is real, we assign the real label to the source-target pair instance. Thus, it is entirely based on the target data label and completely automatic.

(ii). In the second annotation type, we assign the label as" yes" if the source is relevant to the target; otherwise, we assign a label as "no". This annotation is based on the threshold value. To compute the threshold value, we perform Named Entity Recognition (NER) on both source text (S) and target texts (T). The threshold is the ratio of *the number of common entities present in source and target text and the number of entities present in the target text*. We define it as shown in Equation 1, where *R* represents the ratio or threshold. With the help of this threshold value, we find the semantic similarity between source and target text. For this purpose, we make a hypothesis that if the threshold value is greater than 0.5, it may have semantic similarity to some extent. By following this hypothesis, we fix the threshold as 0.5. We assign the label as

[5] https://pypi.org/project/googletrans/

”yes” for the source having a maximum threshold value if it is greater than 0.5. For other sources, we assign the label ”no”.

$$R = \frac{|S \cap T|}{|T|} \tag{1}$$

Human annotation We check the quality of automatic data annotation by performing human annotations for 500 instances of Hindi, Bengali and Tamil languages each. We randomly choose these 500 instances from each language in equal proportionate from fake and real classes. Each instance contains an ID, target-image-URL, target-text, source-URL, source-text, source-image-URL, and source reliability. We provide the selected instances of Hindi and Bengali to native Hindi and Bengali speakers who are proficient in reading, writing, and speaking. Due to the non-availability of Tamil native speakers, we first translated 500 instances of the Tamil language into English. We then provided these translated instances to three English speakers for the annotation. All three annotators are asked to do the following things: (i). Google the target Image URL and open the image in the browser; (ii). Read source-text and find that (a). Source text is related to the image or gives some description of the target image; (b). Source text gives any background information about the target image. If any one of the above points ((a) and (b)) is true, assign the label as ”yes”; otherwise, assign the label as ”no”.

We compute the agreement between the automatic and all three human annotations for the 500 instances of each language using Cohen’s Kappa coefficient (Cohen, 1960). On average, our dataset has 91.27%, 89.5% and 86.3% agreement on Hindi, Bengali and Tamil languages, respectively, indicating a high automatic data annotation quality.

3.3 Data Statistics

In order to create the *MMM* dataset, data instances were collected from Hindi, Bengali, and Tamil languages. We propose a corpus of 10,473 samples having 5630 real and 4840 fake samples. To build the train and test sets, we split the data in an 80:20 ratio. Table 1 outlines the complete data statistics and distribution of MMM dataset. The dataset is organized in a structured way inside the main folder ’Data’ to make them more accessible to researchers. Inside this data directory, there are four folders *viz.* Source, Source Image, Target, and Target Image, and all these 4 folders have 3 sub directories: Hindi, Bengali, and Tamil. The source folder subdirectories contain CSV files corresponding to the language of the source information. All CSV files include attributes such as ID, Number of sources, Source URL, Source text, Image URL, and Reliability. Source Image folder sub-directories contain the source images corresponding to the source language. Target folder sub-directories contain the CSV files corresponding to the language of target information and contain information about the target instance, such as ID, Target URL, Target text, Image URL, and Label. Target Folder subdirectories contain the target image.

Dataset	Total	Real	Fake
Hindi	7163	3563	3600
Bengali	1543	1005	538
Tamil	1767	1065	702
MMM	10473	5633	4840

Table 1: *MMM* dataset statistics and distribution

4 Proposed Model

In this section, we present a brief description of the proposed framework. The overall model is shown in Figure 1 which consists of three components: *Novelty Detection, Image Emotion Prediction, and Misinformation Detection*. Below, we discuss all these three components in details.

4.1 Novelty Detection

We perform a novelty detection task using SCL to find high-level semantic interaction within target and source multimodal news pairs and extract the novelty-aware multimodal feature representations from these news pairs. As discussed below, we give the multimodal source and target as input to the model. We encode the text data using pre-trained MultilingualBERT model (Devlin et al., 2018) and extract the 768-dimensional textual feature representations. To encode the visual data, we use ResNet18 (He et al., 2016) and concatenate the textual and visual features to obtain the multimodal feature representations. We employ two fully connected layers over the encoded source and target representations to project them in a 128-dimensional latent space. Now, we train the model using contrastive learning so that the target representation attracts the source representation if both are of the same class; otherwise, the target repeals the source representation. We optimize the contrastive loss function, similar to Khosla et al. (2020)

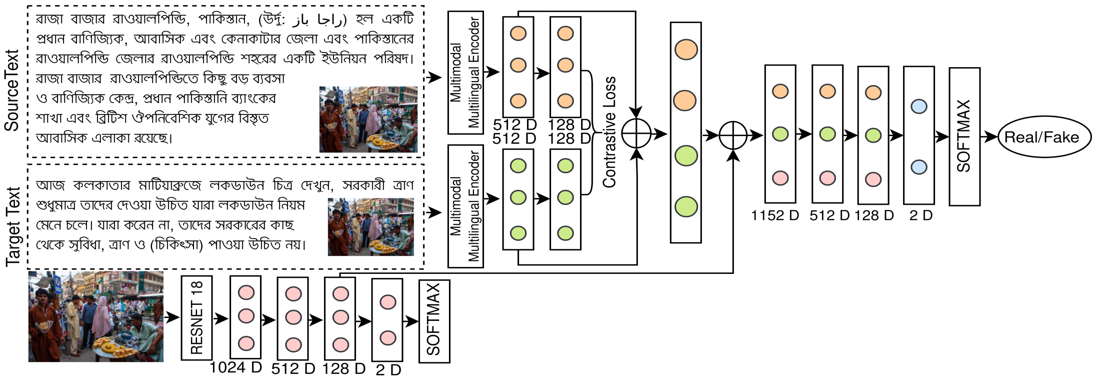

SourceText
রাজা বাজার রাওয়ালপিন্ডি, পাকিস্তান, (উর্দু: راجا باز) হল একটি প্রধান বাণিজ্যিক, আবাসিক এবং কেনাকাটার জেলা এবং পাকিস্তানের রাওয়ালপিন্ডি জেলার রাওয়ালপিন্ডি শহরের একটি ইউনিয়ন পরিষদ। রাজা বাজার রাওয়ালপিন্ডিতে কিছু বড় ব্যবসা ও বাণিজ্যিক কেন্দ্র, প্রধান পাকিস্তানি ব্যাংকের শাখা এবং ব্রিটিশ ঔপনিবেশিক যুগের বিস্তৃত আবাসিক এলাকা রয়েছে।
Target Text
আজ কলকাতার মাটিয়াব্রুজে লকডাউন চিত্র দেখুন, সরকারী ত্রাণ শুধুমাত্র তাদের দেওয়া উচিত যারা লকডাউন নিয়ম মেনে চলে। যারা করেন না, তাদের সরকারের কাছ থেকে সুবিধা, ত্রাণ ও (চিকিৎসা) পাওয়া উচিত নয়।
Multimodal Multilingual Encoder
Multimodal Multilingual Encoder
512 D
512 D
128 D
128 D
Contrastive Loss
1152 D
512 D
128 D
2 D
SOFTMAX
Real/Fake
RESNET 18
1024 D
512 D
128 D
2 D
SOFTMAX

Figure 1: Proposed multilingual multimodal misinformation detection model

to train the novelty model. We mathematically define the loss function in Equation 2. Here, I is the set of indices of the target (anchor); P is the set of positive samples (samples of the same class of anchor), τ is a scalar parameter.

$$L_{SCL} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp(\frac{z_i \cdot z_p}{\tau})}{\sum_{a \in A(i)} exp(\frac{z_i \cdot z_p}{\tau})} \tag{2}$$

4.2 Image Emotion Prediction

Emotional appeal in the news content plays an inevitable role for the spread of false information. A number of prior works on misinformation detection have investigated textual emotions but the visual emotion is still under-explored. In the era of multimedia information, visual emotion convince people to believe in false information much compared to the textual emotion. Motivated by this, we design a neural network-based visual emotion classification model to obtain the emotion-aware visual feature representation. For pre-training this network, we use the combined form of UnbiasedEmo (Panda et al., 2018) and ArtPhoto (Machajdik and Hanbury, 2010) datasets. The instances of the combined datasets are associated with six emotion labels *viz.* joy, love, sadness, fear, surprise, and anger. The study presented by MIT scholars has proved that false rumors usually inspire replies expressing greater surprise, fear and disgust. On the other hand, the true stories inspire greater sadness, anticipation, joy, and trust. Motivated by this investigation, we have kept surprise, fear and disgust in one group and sadness, anticipation, joy, and trust in another group. Categorizing these emotions reflects the news characteristics, which shorten the decision boundary. The dataset that we are using for visual emotion prediction is also highly imbalanced. This is also a major reason for grouping the emotion instances into binary classes. For our experiments, we follow Kumari et al. (2021a) to consider two emotion labels, *emotion true* which is formed by combining joy, love, and sadness labels; and *emotion false* which is formed by the combination of fear, surprise, and anger. Given a set of n images I = (I1, .., In), and their emotion labels EL = (EL1, .., ELn), we encode each image ELi using ResNet18 to the model and pass this encoded image representation through a Multilayer Perceptron (MLP) network that consists of two hidden layers with 1024 and 512 neurons and one output layer with two neurons and a softmax classifier function. Since the number of instances in each emotion class is not balanced, we optimize the weighted cross-entropy loss during training. After training this emotion model, we predict the emotion labels of images present in the developed dataset.

4.3 Misinformation Detection

After pre-training the novelty model, we extract the 512-dimensional feature representations for the source and target then concatenate them to obtain multimodal representation. We project this fused representation into 512-dimensional feature space and use it as a novelty-aware multimodal feature representation to develop our fake news detection model. We also extract 512-dimensional emotion-aware visual feature representations from a pre-trained image emotion model. At last, we concatenate novelty and emotion-aware representations.

After obtaining novelty-aware multimodal representation and emotion-aware visual representation, we concatenate and pass them to MLP that contains two hidden layers and an output layer with a softmax function to classify the news as fake or real. We optimize the cross-entropy loss to train our fake news detection model.

5 Experiments and Results

This section presents experimental setup, baseline, results, case studies and error analysis.

5.1 Experimental Setup

We perform all the experiments with one NVIDIA GeForce RTX GPU and 11GB of RAM using the Pytorch library. We train the baseline models for 100 epochs using the Adam optimizer with 128 batch size. We pre-train the contrastive learning framework for 1000 epochs using LARS optimizer for Stochastic Gradient Descent (SGD) with 512 batch size, which takes approximately 10 minutes. The emotion model is pre-trained using Adam optimizer with 128 batches in 10 minutes, with 100 epochs. We train the final proposed model using the Adam optimizer for 100 epochs for a batch size of 128 which takes approximately 15 minutes.

5.2 Baselines and Comparing Systems

We design some baseline models for validating the performance of our proposed model. We show the results of the proposed and baseline models in Table 3. Apart from these, we also implement the state-of-the-art systems like MLBViT and EANN for the comparison where we feed target text and target image in multimodal feature extractor and use MultilingualBert in place of Text-CNN. We show the results of these comparing systems in Table 2.

Model	Hindi		Bengali		Tamil		MMM	
	FS	Acc	FS	Acc	FS	Acc	FS	Acc
MLBViT	.723	.735	.748	.758	.743	.752	.775	.780
EANN	.833	.822	.845	.856	.870	.883	.855	.868
MVEN	**.939**	**.938**	**.946**	**.945**	**.946**	**.946**	**.955**	**.956**

Table 2: Results of comparing systems. Here, MVEN is our proposed *Multilingual + VisualEmo + Novelty* model; MLBViT: *MultiLingualBert + Vision Transformer*

MLBERT+ResNet: We encode textual and visual information of target using pre-trained MultilingualBERT (Devlin et al., 2018) and pre-trained ResNet18 model (He et al., 2016), respectively. We concatenate the textual and visual representations to obtain multimodal representations and pass this target multimodal representation to MLP network that consists of two hidden layers and one output layer with a softmax function.

MLBERT+ResNet (WBG): We encode the textual and visual information for source and target both similar to the previous baseline model. We concatenate source and target multimodal representation and pass it to MLP network that consists of two hidden layers and one output layer with a softmax classifier function. Thus, in this baseline we also consider source information along with the target information.

Unimodal + VisualEmo: In this model, we encode the target text information using Multilingual-BERT and compute target image emotion using the method, similar to proposed model. We pass the textual representation and emotion aware visual representation to MLP network for the final classification.

Multimodal + VisualEmo: In visualEmo model, we compute the visual emotion similar to the previous baseline. We pass this emotion aware visual representation and source multimodal representation to MLP with Softmax classifier function for the final classification.

Multimodal + Novelty: In novelty model, we implement the proposed framework without emotion module. We apply SCL between source and target multimodal representation to compute the novelty aware representation. We only pass the novelty aware multimodal representation to the MLP with Softmax classifier function.

5.3 Results and Discussion

The results of the baseline models and our proposed model are shown in Table 3. We report the result for our developed *MMM* dataset and also for Hindi, Bengali and Tamil language dataset separately. As shown in Table 3, the *Multilingual + ResNet (WBG)* model performs better than the *Multilingual + ResNet* model for all the datasets which show the importance of background knowledge. *Multimodal + VisualEmo* model produces better results than *Multilingual + ResNet* model. In addition, the *Multimodal + VisualEmo* model performs better than the *Unimodal + VisualEmo* model. The above three factors assist us in concluding that background knowledge, emotion, and multimodality effectively help in fake news prediction. In comparison to the background knowledge framework, we obtain a 2.46 accuracy improvement when we use the *Multimodal + Novelty* model. We can therefore prove that our contrastive learning methodology helps to detect fake news. Compared

Model	Dataset	Fake F1	Real F1	Acc	WA	Model	Dataset	Fake F1	Real F1	Acc	WA
MLBERT + ResNet	Hindi	0.668	0.722	0.697	0.695	**MLBERT + ResNet(WBG)**	Hindi	0.876	0.889	0.876	0.883
	Bengali	0.627	0.813	0.751	0.734		Bengali	0.866	0.879	0.867	0.873
	Tamil	0.666	0.812	0.759	0.750		Tamil	0.872	0.885	0.873	0.880
	MMM	0.703	0.765	0.737	0.737		MMM	0.886	0.901	0.886	0.895
Unimodal + VisualEmo	Hindi	0.813	0.837	0.819	0.826	**Multimodal + VisualEmo**	Hindi	0.861	0.874	0.860	0.868
	Bengali	0.791	0.817	0.799	0.805		Bengali	0.840	0.859	0.848	0.851
	Tamil	0.801	0.826	0.807	0.841		Tamil	0.846	0.836	0.833	0.846
	MMM	0.808	0.851	0.834	0.830		MMM	0.857	0.858	0.857	0.857
Multimodal + Novelty	Hindi	0.919	0.928	0.914	0.925	**Multimodal + VisualEmo + Novelty**	Hindi	0.934	0.942	0.938	0.939
	Bengali	0.900	0.917	0.904	0.910		Bengali	0.939	0.950	0.945	0.946
	Tamil	0.902	0.911	0.899	0.905		Tamil	0.940	0.951	0.946	0.946
	MMM	0.907	0.926	0.910	0.920		MMM	**0.949**	**0.960**	**0.956**	**0.955**

Table 3: Results of the proposed model and its ablated versions. Here, *Multimodal* + *VisualEmo* + *Novelty* is our proposed model; F1: F1 score, Acc: Accuracy, MA: Macro Average, WA: Weighted Average.

to the *Multilingual* + *ResNet* baseline model, our final proposed model (*Multimodal* + *VisualEmo* + *Novelty*) achieves 21.77 accuracy improvement. Hence, our final proposed architecture that utilizes novelty and emotion outperforms all of the baselines and produces the most effective results. We also obtain an 8.8 accuracy improvement over the EANN model.

McNemar significance test (Pembury Smith and Ruxton, 2020) is a well-known statistical test to analyze statistical significance of the differences in classifier's performance. In our work, we also want to prove that the proposed model is comparatively significant with a larger margin than the baseline models. Therefore, we use the McNemar significance test to compute the significance difference between our proposed model and EANN model and obtain p-values 7.3*E-3 that are less than the threshold p-value i.e. 0.05 for rejection of the null hypothesis. It shows that our result is significant.

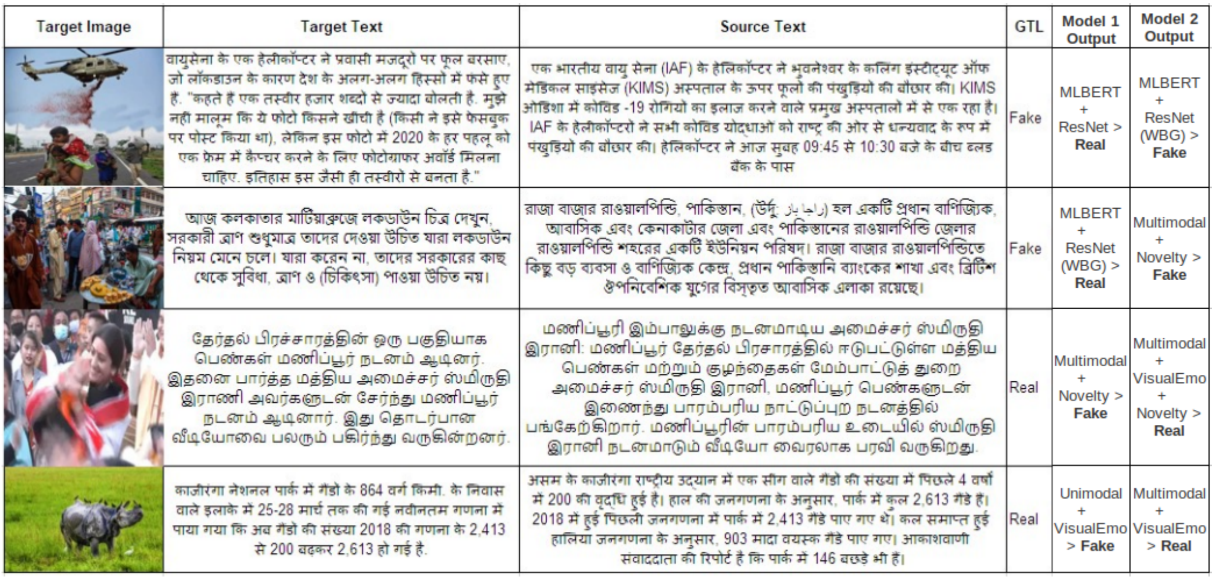
5.4 Case Studies and Error Analysis

We perform a detailed analysis in Figure 2 to show the efficacy of our background knowledge, novelty, emotion and multi-modality. First example shows that concatenation of background Knowledge (source text) with target text help the model to predict accurately. In the second example, source text and target text describe that location of target image is Pakistan and Kolkata, respectively. This mismatch in location is easily detected by *Multimodal* + *Novelty* model which use supervised contrastive learning. In the third example, emotion of target image is joy which is more inclined towards real news so the proposed model with novelty and emotion predicts it accurately. In the last example, visual features of target and source images with source and target text help the model to predict accurately which shows how significant the role multi-modality plays.

We show some examples in Figure 3, which are misclassified by our proposed model. For the first example, the target image shows Avni Chaturvedi, but the target text claims that the image shows Urvisha Jariwala, which is incorrect. *Multimodal* + *Novelty* model focuses solely on novelty and capture the mismatch in source and target text and correctly predicts fake news, but *Multimodal* + *VisualEmo* + *Novelty* model gives the wrong prediction because emotion associated with this image is joy which is an attribute of true news, so it misleads the model. In the second example, the model with novelty and emotion performs better than with model that doesbackground knowledge (WBG) model. Novelty emotion model can flag this news as fake based on the source text collected which clearly states that the original image was an old image and taken in 2014s. However, the mismatch between source and target text is not noticeable with background knowledge model, resulting in incorrect predictions. For the last example, *Multimodal* + *VisualEmo* model performs well than our proposed model. With novelty and emotion, we see that the source text and target text both give some information about covid-19 but the source text has some additional information about the election while the target text gives more emphasis on symptoms which mislead the model and contrastive learning takes it away from the main subject.

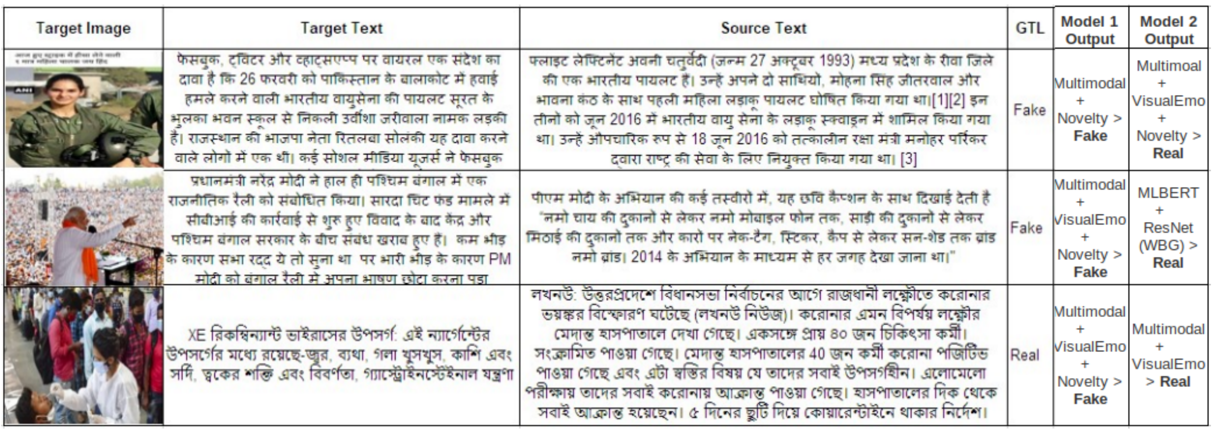
6 Conclusion

In this paper we solve the problem of multilingual multimodal misinformation detection in three Indian languages, Hindi, Bengali and Tamil. Nowa-days, same image is used in different textual context to mislead the reader. To address this problem, first, we have created our Multilingual Multimodal Misinformation dataset and then we



Target Image	Target Text	Source Text	GTL	Model 1 Output	Model 2 Output
	वायुसेना के एक हेलीकॉप्टर ने प्रवासी मजदूरो पर फूल बरसाए, जो लॉकडाउन के कारण देश के अलग-अलग हिस्सो में फंसे हुए हैं. "कहते हैं एक तस्वीर हजार शब्दो से ज्यादा बोलती है. मुझे नहीं मालूम कि ये फोटो किसने खींची है (किसी ने इसे फेसबुक पर पोस्ट किया था), लेकिन इस फोटो में 2020 के हर पहलू को एक फ्रेम में कैप्चर करने के लिए फोटोग्राफर अवॉर्ड मिलना चाहिए. इतिहास इस जैसी ही तस्वीरो से बनता है."	एक भारतीय वायु सेना (IAF) के हेलिकॉप्टर ने भुवनेश्वर के कलिंग इंस्टीट्यूट ऑफ मेडिकल साइंसेज (KIMS) अस्पताल के ऊपर फूलो की पंखुड़ियो की बौछार की। KIMS ओडिशा में कोविड -19 रोगियो का इलाज करने वाले प्रमुख अस्पतालो में से एक रहा है। IAF के हेलीकॉप्टरो ने सभी कोविड योद्धाओं को राष्ट्र की ओर से धन्यवाद के रूप में पंखुड़ियो की बौछार की। हेलिकॉप्टर ने आज सुबह 09:45 से 10:30 बजे के बीच ब्लड बैंक के पास	Fake	MLBERT + ResNet > **Real**	MLBERT + ResNet (WBG) > **Fake**
	আজ কলকাতার মাটিয়াব্রুজে লকডাউন চিত্র দেখুন, সরকারী ত্রাণ শুধুমাত্র তাদের দেওয়া উচিত যারা লকডাউন নিয়ম মেনে চলে। যারা করেন না, তাদের সরকারের কাছ থেকে সুবিধা, ত্রাণ ও (চিকিৎসা) পাওয়া উচিত নয়।	রাজা বাজার রাওয়ালপিন্ডি, পাকিস্তান, (উর্দু: راجا بازار) হল একটি প্রধান বাণিজ্যিক, আবাসিক এবং কেনাকাটার জেলা এবং পাকিস্তানের রাওয়ালপিন্ডি জেলার রাওয়ালপিন্ডি শহরের একটি ইউনিয়ন পরিষদ। রাজা বাজার রাওয়ালপিন্ডিতে কিছু বড় ব্যবসা ও বাণিজ্যিক কেন্দ্র, প্রধান পাকিস্তানি ব্যাংকের শাখা এবং ব্রিটিশ ঔপনিবেশিক যুগের বিস্তৃত আবাসিক এলাকা রয়েছে।	Fake	MLBERT + ResNet (WBG) > **Real**	Multimodal + Novelty > **Fake**
	தேர்தல் பிரச்சாரத்தின் ஒரு பகுதியாக பெண்கள் மணிப்பூர் நடனம் ஆடினர். இதனை பார்த்த மத்திய அமைச்சர் ஸ்மிருதி இராணி அவர்களுடன் சேர்ந்து மணிப்பூர் நடனம் ஆடினார். இது தொடர்பான வீடியோவை பலரும் பகிர்ந்து வருகின்றனர்.	மணிப்பூரி இம்பாலுக்கு நடனமாடிய அமைச்சர் ஸ்மிருதி இரானி: மணிப்பூர் தேர்தல் பிரசாரத்தில் ஈடுபட்டுள்ள மத்திய பெண்கள் மற்றும் குழந்தைகள் மேம்பாட்டுத் துறை அமைச்சர் ஸ்மிருதி இரானி, மணிப்பூர் பெண்களுடன் இணைந்து பாரம்பரிய நாட்டுப்புற நடனத்தில் பங்கேற்கிறார். மணிப்பூரின் பாரம்பரிய உடையில் ஸ்மிருதி இரானி நடனமாடும் வீடியோ வைரலாக பரவி வருகிறது.	Real	Multimodal + Novelty > **Fake**	Multimodal + VisualEmo + Novelty > **Real**
	काजीरंगा नेशनल पार्क में गैंडो के 864 वर्ग किमी. के निवास वाले इलाके में 25-28 मार्च तक की गई नवीनतम गणना में पाया गया कि अब गैंडो की संख्या 2018 की गणना के 2,413 से 200 बढ़कर 2,613 हो गई है.	असम के काजीरंगा राष्ट्रीय उद्यान में एक सींग वाले गैंडो की संख्या में पिछले 4 वर्षो में 200 की वृद्धि हुई है। हाल की जनगणना के अनुसार, पार्क में कुल 2,613 गैंडे हैं। 2018 में हुई पिछली जनगणना में पार्क में 2,413 गैंडे पाए गए थे। कल समाप्त हुई हालिया जनगणना के अनुसार, 903 मादा वयस्क गैंडे पाए गए। आकाशवाणी संवाददाता की रिपोर्ट है कि पार्क में 146 बछड़े भी हैं।	Real	Unimodal + VisualEmo > **Fake**	Multimodal + VisualEmo > **Real**

Figure 2: Some case studies where model 2 correctly classifies the misinformation. Here, GTL: Ground Truth Label, Model 1 Output and Model 2 Output are the different models output shown in that particular column.



Target Image	Target Text	Source Text	GTL	Model 1 Output	Model 2 Output
	फेसबुक, ट्विटर और व्हाट्सएप्प पर वायरल एक संदेश का दावा है कि 26 फरवरी को पाकिस्तान के बालाकोट में हवाई हमले करने वाली भारतीय वायुसेना की पायलट सूरत के भुलका भवन स्कूल से निकली उर्वीशा जरीवाला नामक लड़की है। राजस्थान की भाजपा नेता रितलबा सोलंकी यह दावा करने वाले लोगो में एक थीं। कई सोशल मीडिया यूजर्स ने फेसबुक	फ्लाइट लेफ्टिनेंट अवनी चतुर्वेदी (जन्म 27 अक्टूबर 1993) मध्य प्रदेश के रीवा जिले की एक भारतीय पायलट हैं। उन्हें अपने दो साथियो, मोहना सिंह जीतरवाल और भावना कंठ के साथ पहली महिला लड़ाकू पायलट घोषित किया गया था।[1][2] इन तीनो को जून 2016 में भारतीय वायु सेना के लड़ाकू स्क्वाड्रन में शामिल किया गया था। उन्हें औपचारिक रूप से 18 जून 2016 को तत्कालीन रक्षा मंत्री मनोहर पर्रिकर द्वारा राष्ट्र की सेवा के लिए नियुक्त किया गया था। [3]	Fake	Multimodal + Novelty > **Fake**	Multimoal + VisualEmo + Novelty > **Real**
	प्रधानमंत्री नरेंद्र मोदी ने हाल ही पश्चिम बंगाल में एक राजनीतिक रैली को संबोधित किया। सारदा चिट फंड मामले में सीबीआई की कार्रवाई से शुरु हुए विवाद के बाद केंद्र और पश्चिम बंगाल सरकार के बीच संबंध खराब हुए हैं। कम भीड़ के कारण सभा रद्द ये तो सुना था पर भारी भीड़ के कारण PM मोदी को बंगाल रैली मे अपना भाषण छोटा करना पड़ा	पीएम मोदी के अभियान की कई तस्वीरो में, यह छवि कैप्शन के साथ दिखाई देती है "नमो चाय की दुकानो से लेकर नमो मोबाइल फोन तक, साड़ी की दुकानो से लेकर मिठाई की दुकानो तक और कारो पर नेक-टैग, स्टिकर, कैप से लेकर सन-शेड तक ब्रांड नमो ब्रांड। 2014 के अभियान के माध्यम से हर जगह देखा जाना था।"	Fake	Multimodal + VisualEmo + Novelty > **Fake**	MLBERT + ResNet (WBG) > **Real**
	XE রিকম্বিন্যান্ট ভাইরাসের উপসর্গ: এই ন্যার্গেন্টের উপসর্গের মধ্যে রয়েছে-জ্বর, ব্যথা, গলা খুসখুস, কাশি এবং সর্দি, ত্বকের শক্তি এবং বিবর্ণতা, গ্যাস্ট্রোইনস্টেইনাল যন্ত্রণা	লখনউ: উত্তরপ্রদেশে বিধানসভা নির্বাচনের আগে রাজধানী লক্ষ্ণৌতে করোনার ভয়ঙ্কর বিস্ফোরণ ঘটেছে (লখনউ নিউজ)। করোনার এমন বিপর্যয় লক্ষ্ণৌর মেদান্ত হাসপাতালে দেখা গেছে। একসঙ্গে প্রায় ৪০ জন চিকিৎসা কর্মী। সংক্রামিত পাওয়া গেছে। মেদান্ত হাসপাতালের 40 জন কর্মী করোনা পজিটিভ পাওয়া গেছে এবং এটা স্বস্তির বিষয় যে তাদের সবাই উপসর্গহীন। এলোমেলো পরীক্ষায় তাদের সবাই করোনায় আক্রান্ত পাওয়া গেছে। হাসপাতালের দিক থেকে সবাই আক্রান্ত হয়েছেন। ৫ দিনের ছুটি দিয়ে কোয়ারেন্টাইনে থাকার নির্দেশ।	Real	Multimodal + VisualEmo + Novelty > **Fake**	Multimodal + VisualEmo > **Real**

Figure 3: Error analysis on some examples which are misclassified.

have performed experiments on Multilinguality, Background knowledge, Emotion, Multimodality and Novelty to see the effect. We have built a novel framework based on novelty and emotion which outperform all the baseline and state-of-the-art models. Further, We want to extend our current work in following direction to prevent the spread of misinformation: (i). by including additional low-resources language; (ii). by addressing cross-lingual and code-mixed based resources and mechanisms; and (iii). by incorporating explainability in the model.

Ethical Declaration

We have developed our dataset using publicly available information on different websites. Our use of those data has adhered to the policy guidelines and has not caused any copyright issues. During the creation of our dataset, we collected news articles and related information that did not contain any sensitive information. We will make the data available only for research purposes after signing an agreement.

Acknowledgments

This reserach is partially supported by the project “HELIOS - Hate, Hyperpartisan, and Hyperpluralism Elicitation and Observer System“, sponsored by Wipro. We thank all the anonymous reviewers for their insightful suggestions. We also sincerely thank Tanik Saikh for his contribution in data annotation of Bengali Language. All the research work was carried out at AI-ML-NLP research lab, IIT Patna.

References

Hugo Queiroz Abonizio, Janaina Ignacio de Morais, Gabriel Marques Tavares, and Sylvio Barbon Junior. 2020. Language-independent fake news detection: English, portuguese, and spanish mutual features. *Future Internet*, 12(5):87.

Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, Yiannis Kompatsiaris, et al. 2015. Verifying multimedia use at mediaeval 2015. *MediaEval*, 3(3):7.

Ali K Chaudhry, Darren Baker, and Philipp Thun-Hohenstein. 2017. Stance detection for the fake news challenge: identifying textual relationships with deep neural nets. *CS224n: Natural Language Processing with Deep Learning*.

M Chen, X Chu, and KP Subbalakshmi. 2021. Mmcovar: Multimodal covid-19 vaccine focused data repository for fake news detection and a baseline architecture for classification.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Maria Glenski, Ellyn Ayton, Josh Mendoza, and Svitlana Volkova. 2019. Multilingual multimodal digital deception detection and disinformation spread across social platforms. *arXiv preprint arXiv:1909.05838*.

Gaël Guibon, Liana Ermakova, Hosni Seffih, Anton Firsov, and Guillaume Le Noé-Bienvenu. 2019. Multilingual fake news detection with satire. In *CICLing: International Conference on Computational Linguistics and Intelligent Text Processing*.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2021. Arcov19-rumors: Arabic covid-19 twitter dataset for misinformation detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 72–81.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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*, pages 795–816.

Sarthak Jindal, Raghav Sood, Richa Singh, Mayank Vatsa, and Tanmoy Chakraborty. 2020. Newsbag: A multimodal benchmark dataset for fake news detection.

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*, pages 2915–2921.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33.

Rina Kumari, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. 2021a. Misinformation detection using multitask learning with mutual learning for novelty detection and emotion recognition. *Information Processing & Management*, 58(5):102631.

Rina Kumari, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. 2021b. A multitask learning approach for fake news detection: Novelty, emotion, and sentiment lend a helping hand. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Rina Kumari, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. 2022. What the fake? probing misinformation detection standing on the shoulder of novelty and emotion. *Information Processing & Management*, 59(1):102740.

Rina Kumari and Asif Ekbal. 2021. Amfb: Attention based multimodal factorized bilinear pooling for multimodal fake news detection. *Expert Systems with Applications*, 184:115412.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020a. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *arXiv preprint arXiv:2011.04088*.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020b. Toward a multilingual and multimodal data repository for covid-19 disinformation. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4325–4330. IEEE.

Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6149–6157.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of* *the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

Dan Saattrup Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. *arXiv preprint arXiv:2202.11684*.

Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. 2018. Contemplating visual emotions: Understanding and overcoming dataset bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 579–595.

Matilda QR Pembury Smith and Graeme D Ruxton. 2020. Effective use of the mcnemar test. *Behavioral Ecology and Sociobiology*, 74(11):1–9.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Tanik Saikh, Arkadipta De, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A deep learning approach for automatic detection of fake news. *arXiv preprint arXiv:2005.04938*.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Defend: Explainable fake news detection. In *25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2019*, pages 395–405. Association for Computing Machinery.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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

Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2021. Factorization of fact-checks for low resource indian languages. *arXiv preprint arXiv:2102.11276*.

Valeriya Slovikovskaya and Giuseppe Attardi. 2020. Transfer learning from transformers to fake news challenge stance detection (fnc-1) task. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1211–1218.

Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2021. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management*, 58(1):102437.

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 & data mining*, pages 849–857.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2560–2569.

Chen Yang, Xinyi Zhou, and Reza Zafarani. 2021. Checked: Chinese covid-19 fake news dataset. *Social Network Analysis and Mining*, 11(1):1–8.

Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S Yu. 2018. Ti-cnn: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749*.

Wenpeng Yin and Dan Roth. 2018. Twowingos: A two-wing optimization strategy for evidential claim verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114.

Jiawei Zhang, Bowen Dong, and S Yu Philip. 2020. Fakedetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1826–1829. IEEE.

Wenjia Zhang, Lin Gui, and Yulan He. 2021. Supervised contrastive learning for multimodal unreliable news detection in covid-19 pandemic. *arXiv preprint arXiv:2109.01850*.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3205–3212.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108.

A Data Statistics

We have also computed the average and median length of the source and target text for Hindi, Bengali, Tamil and MMM datasets shown in Table 4. The average target text length of the Tamil language is smaller than other languages.

B Data Collection Flow Diagram

Figure 4 shows the complete data preparation flow diagram we have followed to collect background knowledge. We collect the target instances of our *MMM* dataset using the following flow diagram:

C Multilingual analysis

We also tested the trained model on language which is not included in the training process, as a zero-shot experiment. For this, we make three groups of dataset *Hindi+Bengali*, *Bengali+Tamil*, *Tamil+Hindi* and train the model on each group. Finally, we test the model on different combination of unseen and seen language. This experiment shows that the model can be generalized for an unseen language also by using language-independent features.

1.) Firstly, we test the model on Tamil, Hindi, and Bengali languages, which the model does not see during training. This experiment shows that the model can also be generalized for an unseen language by using language-independent features. The first section of Table 5 *Multilingual training with Monolingual testing on unseen language* shows the model's performance is the least when it is evaluated on the test set of Tamil data. This is because Hindi and Bengali belong to the same language family, i.e., the Indo-Aryan language family. In contrast, Tamil belongs to the Dravidian language family, resulting in less generalization of the model.

2.) We also test the model with test data, having all three language. It means this time; we consider both seen language and unseen language. The second section of Table 5 *Multilingual training with Multilingual testing on seen and unseen language* shows that model is performing slightly better than the first section of table 5 because training data include seen language also.

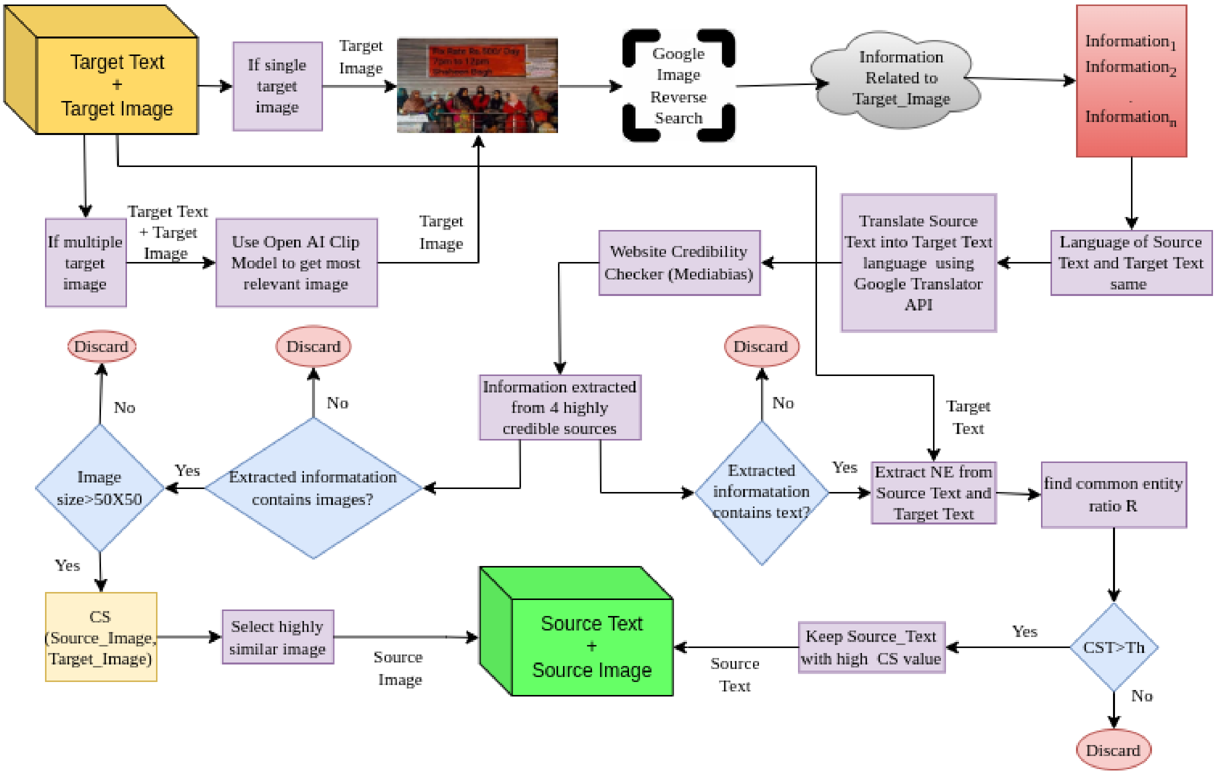
3.) The third section of Table 5 depicts the results for *Multilingual training with Monolingual testing on seen language*. Here, we train the model in two-step *viz.* i). We train the model with three language groups having two languages in each group and train the model with each group's language, respectively and ii). we train the model with all three languages and feed monolingual test data for all three languages.

D Translated version of case studies and error analysis

We have also translated case studies and error analysis into English language in Figure 5 and Figure 6 respectively so that everyone can understand it.

Dataset	Length	Target	Source_1	Source_2	Source_3	Source_4
Hindi	Average	103.22	1668.44	779.52	637.48	586.46
	Median	26	805.5	576	540	479
Bengali	Average	68.69	1964.26	584	471.07	403.81
	Median	12	772	396	367	302
Tamil	Average	37.43	1298.07	518.2	487.31	380.61
	Median	8	512	281	268	217
MMM	Average	87.93	1650.59	710.7	590.41	528.27
	Median	22	743.5	509	469	402

Table 4: Dataset statistics on Source and Target text length



Target Text + Target Image
If single target image
Target Image
Google Image Reverse Search
Information Related to Target_Image
Information$_1$
Information$_2$
Information$_n$
If multiple target image
Target Text + Target Image
Use Open AI Clip Model to get most relevant image
Target Image
Website Credibility Checker (Mediabias)
Translate Source Text into Target Text language using Google Translator API
Language of Source Text and Target Text same
Discard
Discard
Discard
No
No
No
Information extracted from 4 highly credible sources
Target Text
Image size>50X50
Yes
Extracted informatation contains images?
Extracted information contains text?
Yes
Extract NE from Source Text and Target Text
find common entity ratio R
Yes
CS (Source_Image, Target_Image)
Select highly similar image
Source Image
Source Text + Source Image
Source Text
Keep Source_Text with high CS value
Yes
CST>Th
No
Discard

Figure 4: Flow diagram to collect source information from each target instance

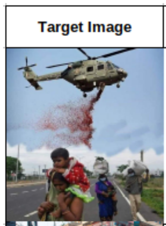
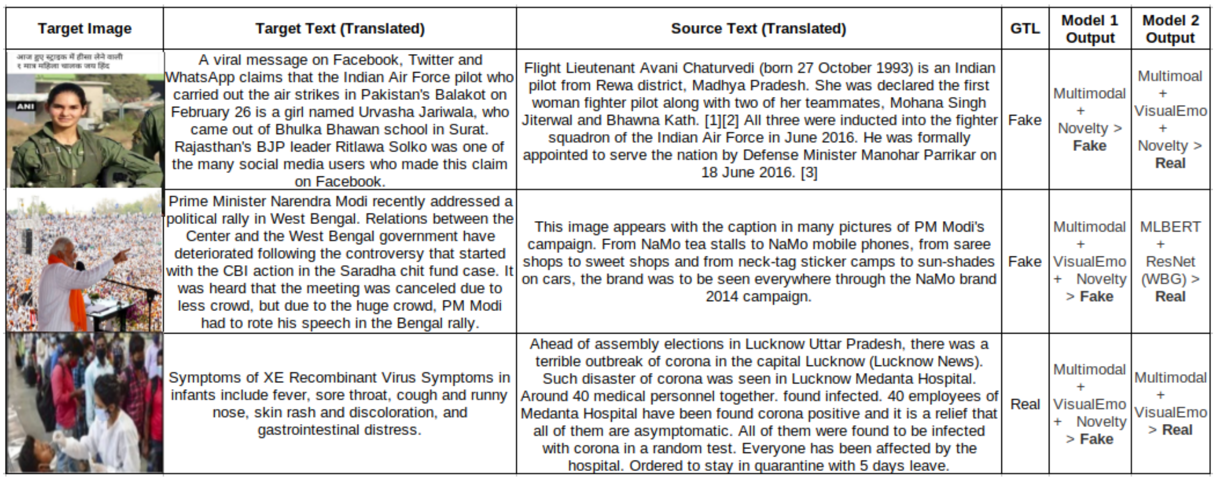
Target Image	Target Text (Translated)	Source Text (Translated)	GTL	Model 1 Output	Model 2 Output
	An Air Force helicopter showered flowers on migrant laborers who have become firm in different parts of the country due to the lockdown. It is said that a picture speaks more than a thousand words, I don't know who took this photo (somebody posted it on Facebook, but this photo should get the Photographer Award for capturing every aspect of 2020 in one frame) History is made from such pictures.	An Indian Air Force (IAF) helicopter showers flower petals over the Kalang Institute of Medical Sciences (KIMS) Hospital in Bhubaneswar. KIMS has been one of the major hospitals in Odisha treating COVID 19 patients. IAF helicopters showered vertical showers of thanks on behalf of the nation to all the covid warriors. Helicopters from 09.45 am to 10:30 am today near Lad Bank	Fake	MLBERT + ResNet > **Real**	MLBERT + ResNet (WBG) > **Fake**
	View Lockdown Images Today in Martiaktje, Kolkata Government knowledge should be given only to those who follow the lockdown rules. Those who do not, should not get benefits, food and (treatment) from the government.	Raja Bazar Rawalpindi Pakistan, (Urdu 1961) is a major commercial residential and a union council of Rawalpindi city in Kenakartar district and Rawalpindi district of Pakistan. Raja Bazar in Rawalpindi Some major business and commercial centers include branches of major Pakistani banks and extensive residential areas dating back to the British colonial era.	Fake	MLBERT + ResNet (WBG) > **Real**	Multimodal + Novelty > **Fake**
	Women perform Manipur dance as part of election campaign. Seeing this Union Minister Samruti Irani danced Manipur along with them. Many people are sharing the related video.	Minister Smriti danced to Manipuri Imphal Irani Manipur Election Campaign Union Women and Child Development Minister Smriti Irani participates in a traditional folk dance with Manipur women. A video of Smriti Irani dancing in traditional Manipur attire is going viral.	Real	Multimodal + Novelty > **Fake**	Multimodal + VisualEmo + Novelty > **Real**
	In Kaziranga National Park, the latest census of 864 sq km inhabited by pits from March 25-28 found that the number of pits has now increased by 200 to 2,613 from 2,413 in the 2018 census.	In Assam's Jarga National Park, the number of one-sug-shaped pits has increased by 200 in the last 4 years. According to the most recent census, there are a total of 2,613 pits in the park. In the last census conducted in 2018, 2.413 rhinos were found in the park. According to the most recent census that ended yesterday, 903 female adult litters were found. All India Radio reporter reports that there are also 146 calves in the park.	Real	Unimodal + VisualEmo > **Fake**	Multimodal + VisualEmo > **Real**

Figure 5: Some case studies where model 2 correctly classifies the misinformation. Here, GTL: Ground Truth Label, Model 1 Output and Model 2 Output are the different models output shown in that particular column.

Train	Test	Fake			Real			Acc	MA	WA
		P	R	F1	P	R	F1			
Multilingual training with Monolingual testing on unseen language										
H+B	T	0.880	0.871	0.875	0.893	0.900	0.896	0.894	0.886	0.885
B+T	H	0.892	0.889	0.891	0.907	0.910	0.908	0.918	0.899	0.903
T+H	B	0.878	0.868	0.873	0.890	0.899	0.894	0.906	0.884	0.883
Multilingual training with Multilingual testing on seen and unseen language										
H+B	H+B+T	0.9305	0.915	0.920	0.925	0.942	0.934	0.927	0.927	0.927
B+T	H+B+T	0.935	0.904	0.919	0.922	0.941	0.931	0.926	0.925	0.926
T+H	H+B+T	0.9297	0.899	0.914	0.917	0.942	0.930	0.923	0.922	0.923
Multilingual training with Monolingual testing on seen language										
H+B	H	0.937	0.941	0.939	0.950	0.946	0.948	0.944	0.944	0.944
H+B	B	0.936	0.931	0.933	0.942	0.935	0.943	0.939	0.938	0.939
B+T	B	0.935	0.919	0.925	0.944	0.968	0.956	0.942	0.940	0.941
B+T	T	0.931	0.920	0.925	0.933	0.942	0.938	0.932	0.932	0.932
T+H	H	0.936	0.925	0.931	0.938	0.947	0.942	0.937	0.936	0.937
T+H	T	0.930	0.915	0.923	0.929	0.936	0.933	0.930	0.928	0.930
H+B+T	H	0.951	0.931	0.941	0.957	0.972	0.964	0.954	0.953	0.952
H+B+T	B	0.952	0.936	0.944	0.947	0.957	0.952	0.949	0.948	0.948
H+B+T	T	0.9526	0.946	0.949	0.955	0.962	0.958	0.954	0.953	0.954

Table 5: Results on a different combination of training and testing language; Here P, R, F-S are Precision, Recall and F1 score, respectively; Acc: Accuracy, MA: Macro Average, WA: Weighted Average; H:Hindi, B:Bengali, T:Tamil


आज हुए स्ट्राइक में हीसा लेने वाली
१ मात्र महिला चालक जय हिंद
ANI

Target Image	Target Text (Translated)	Source Text (Translated)	GTL	Model 1 Output	Model 2 Output
	A viral message on Facebook, Twitter and WhatsApp claims that the Indian Air Force pilot who carried out the air strikes in Pakistan's Balakot on February 26 is a girl named Urvasha Jariwala, who came out of Bhulka Bhawan school in Surat. Rajasthan's BJP leader Ritlawa Solko was one of the many social media users who made this claim on Facebook.	Flight Lieutenant Avani Chaturvedi (born 27 October 1993) is an Indian pilot from Rewa district, Madhya Pradesh. She was declared the first woman fighter pilot along with two of her teammates, Mohana Singh Jiterwal and Bhawna Kath. [1][2] All three were inducted into the fighter squadron of the Indian Air Force in June 2016. He was formally appointed to serve the nation by Defense Minister Manohar Parrikar on 18 June 2016. [3]	Fake	Multimodal + Novelty > **Fake**	Multimoal + VisualEmo + Novelty > **Real**
	Prime Minister Narendra Modi recently addressed a political rally in West Bengal. Relations between the Center and the West Bengal government have deteriorated following the controversy that started with the CBI action in the Saradha chit fund case. It was heard that the meeting was canceled due to less crowd, but due to the huge crowd, PM Modi had to rote his speech in the Bengal rally.	This image appears with the caption in many pictures of PM Modi's campaign. From NaMo tea stalls to NaMo mobile phones, from saree shops to sweet shops and from neck-tag sticker camps to sun-shades on cars, the brand was to be seen everywhere through the NaMo brand 2014 campaign.	Fake	Multimodal + VisualEmo + Novelty > **Fake**	MLBERT + ResNet (WBG) > **Real**
	Symptoms of XE Recombinant Virus Symptoms in infants include fever, sore throat, cough and runny nose, skin rash and discoloration, and gastrointestinal distress.	Ahead of assembly elections in Lucknow Uttar Pradesh, there was a terrible outbreak of corona in the capital Lucknow (Lucknow News). Such disaster of corona was seen in Lucknow Medanta Hospital. Around 40 medical personnel together. found infected. 40 employees of Medanta Hospital have been found corona positive and it is a relief that all of them are asymptomatic. All of them were found to be infected with corona in a random test. Everyone has been affected by the hospital. Ordered to stay in quarantine with 5 days leave.	Real	Multimodal + VisualEmo + Novelty > **Fake**	Multimodal + VisualEmo > **Real**

Figure 6: Error analysis on some examples which are misclassified.