

Prediction of People’s Emotional Response towards Multi-modal News

Ge Gao, Sejin Paik, Carley Reardon, Yanling Zhao, Lei Guo,
Prakash Ishwar, Margrit Betke, Derry Wijaya

Boston University

{ggao02, sejin, reardonc, lingzhao, guolei,
pi, betke, wijaya}@bu.edu

Abstract

We aim to develop methods for understanding how multimedia news exposure can affect people’s emotional responses, and we especially focus on news content related to gun violence, a very important yet polarizing issue in the U.S. We created the dataset NEmo⁺ by significantly extending the U.S. gun violence news-to-emotions dataset, BU-NEmo, from 320 to 1,297 news headline and lead image pairings and collecting 38,910 annotations in a large crowdsourcing experiment. In curating the NEmo⁺ dataset, we developed methods to identify news items that will trigger similar versus divergent emotional responses. For news items that trigger similar emotional responses, we compiled them into the NEmo⁺-Consensus dataset. We benchmark models on this dataset that predict a person’s *dominant* emotional response toward the target news item (single-label prediction). On the full NEmo⁺ dataset, containing news items that would lead to both differing and similar emotional responses, we also benchmark models for the novel task of predicting the *distribution* of evoked emotional responses in humans when presented with multi-modal news content. Our single-label and multi-label prediction models outperform baselines by large margins across several metrics.

1 Introduction

Understanding how exposure to certain textual and visual news affects people’s emotional reactions is important for detecting, educating, and correcting intentional or unintentional emotional manipulation of readers. As a step towards detecting such manipulations and raising news consumers’ visual literacy, in this work we develop methods for predicting emotional responses towards news headlines and images. To the best of our knowledge, machine learning tools that predict how a reader will react emotionally to a certain news headline, choice of a lead image, or combination of both do

not exist. In this paper, we introduce tools that enable such prediction and thus can shed light on effects of news presentation, which is important to both editors and consumers of news.

The dataset we utilize in this work has been developed in phases. It first started with the headlines of news articles in the Gun Violence Frame Corpus (GVFC) (Liu et al., 2019), along with corresponding lead images of these articles (Tourni et al., 2021). A previous study started a crowd-sourcing experiment to collect emotional response annotations to the news headlines and images, producing the BU-NEmo dataset (Reardon et al., 2022). In this work, we extend the above emotional response experiment significantly. We utilize our new expanded dataset, named NEmo⁺, and present the first benchmark of models to predict the evoked emotional responses in news consumers when presented with multi-modal news content.

2 Related Works

2.1 Predicting Emotional Responses to Text

Sentiment analysis is the task of detecting positive vs. negative sentiment expressed by text. The previous works on text-based emotion prediction have mostly focused on binary classification of positive versus negative emotions (Jiang et al., 2011; Wang et al., 2018). In our work, we aim to predict which category of emotions, from multiple choices, a text will elicit, a task for which there is limited prior work. Ahmad et al. (2020) focus on multi-class emotion state classification in poetry and Vasava et al. (2022) aimed to predict the type of emotion in essays written in response to newspaper articles. While Vasava et al. (2022) classified each essay into one of six basic emotions (Ekman and Friesen, 1971), we use the eight emotions from the prominent psychological study by Mikels et al. (2005) as our categories. The major difference between our work and that of Vasava et al. (2022) is that the essays used in their study already con-

tain readers’ sentiments on the newspaper articles. We present the novel task of directly predicting the emotional reactions of readers to news headline text, without such essays. The recent study of [Gabriel et al. \(2022\)](#) involves modeling how readers react to news headlines. Their work however, focuses on free-text explanations of readers’ reactions and ordinal estimates of likelihood of spread and identification of real vs fake news headlines. By contrast, our dataset contains categorical emotional labels in order to predict the emotional responses. [Gambino and Calvo \(2019\)](#)’s study is closely relevant to ours as they also focused on the novel task of predicting the evoked emotion rather than the previous research of identifying the presence or absence of an emotion. They collected a group of news articles and their associated tweet responses and annotated the emotions expressed in them. They are predicting the evoked emotions towards the whole news article and we are using only the headline as we aim to explore how specific choices of the headline text by the news editors affect the emotion reactions.

2.2 Predicting Emotional Responses to Images

Recent computer vision work has focused on building models to recognize the emotional state of specific persons in images ([Li et al., 2021](#); [Zhang and Xu, 2022](#)), rather than the emotional state that images can elicit in humans. There is very limited work on predicting these reactions to visual data ([Machajdik and Hanbury, 2010](#); [You et al., 2016](#); [Achlioptas et al., 2021](#)). The most relevant of these works is the ArtEmis dataset ([Achlioptas et al., 2021](#)), which contains more than 80k art-related images with annotations of (1) emotional reactions of crowdworkers towards images and (2) their free-flowing English textual explanations of how and why they felt a certain way. Studies with ArtEmis predict (1) by analyzing (2), a task far simpler than ours since their model input is an explanation of an emotion that the model then learns to extract. In our task, emotional reactions must be predicted from the original news headlines and images.

2.3 Predicting Emotional Responses to Multi-modal Content

Multi-modal models have gained success in predicting and understanding emotions by combining audio, textual, and visual data ([Busso et al., 2008](#); [Poria et al., 2019](#); [Dudzik et al., 2020](#)). Most of the previous multi-modal models for emotional pre-

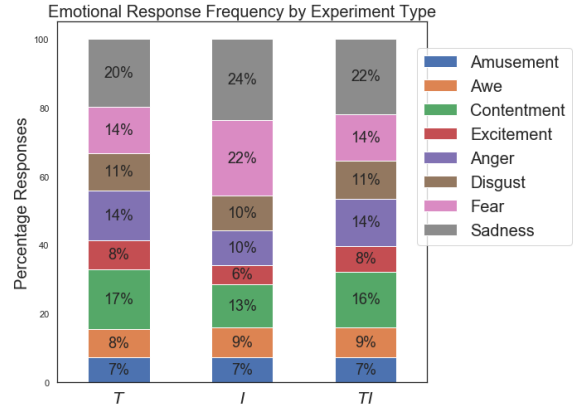


Figure 1: Distributions of emotional responses in the NEMO⁺ dataset by experimental condition (T, I, TI). Evidently, given the nature of gun violence news, the annotated emotions are imbalanced and have an inclination towards negative emotions like sadness and fear.

diction focused on combining elements that are homologous in nature. For example, the MELD dataset ([Poria et al., 2019](#)) predicts emotions using multiple modalities (audio, textual, and visual), which were all part of the same video source. The BU-NEMO dataset created by [Reardon et al. \(2022\)](#) is novel in that the modalities (news headline and image) were separate in nature and chosen to be presented together by the news publishers. We significantly extended this dataset to create the NEMO⁺ dataset in order to have enough training data for multi-modal models. Multi-modal learning on the NEMO⁺ dataset can give us an idea of the likely emotional reactions evoked by a specific combination of inputs from multiple modalities (news headlines and images).

There are limited datasets available for pre-training our models for both text-to-emotion and image-to-emotion prediction. Most of the datasets mentioned above use different sets of emotional labels than in our NEMO⁺ dataset. ArtEmis ([Achlioptas et al., 2021](#)) provides the same 8 emotional labels as ours in addition to a 9th "something else," so we used ArtEmis to pre-train some of our models.

3 Data

3.1 Dataset Collection

BU-NEMO ([Reardon et al., 2022](#)) previously extended the work of the Gun Violence Frame Corpus (GVFC) ([Liu et al., 2019](#); [Guo et al., 2021](#)) which applied frame detection on gun violence related news headlines, and created 1,300 news headline and image pairings. [Reardon et al. \(2022\)](#) initially annotated the news items in GVFC with emo-

tional responses by workers from Amazon Mechanical Turk (MTurk) with annotators of at least high school qualification. A significant portion of the annotations contained spam in the free flow written text making the quality of the categorical emotional responses questionable. This spamming on MTurk is consistent with other findings of MTurk’s low annotation quality (Rashtchian et al., 2010). Due to this limitation, we decided to implement a survey website (hosted on AWS) with the same survey content and interface to the study of Reardon et al. (2022) to collect the annotations for this study. We awarded course credits to anonymous student participants from the College of Communication and the Computer Science Department at Boston University through an internal annotation collection system managed by the university. We received high quality responses.

For our data collection, we followed the same pipeline as the BU-NEMO study (Reardon et al., 2022). Our pool of annotators consisted of undergraduate and graduate university students. The BU-NEMO dataset contained 320 news items with 10,547 annotations. Our NEMO⁺ dataset is significantly expanded, by adding 977 news items and 28,363 annotations to the original dataset. For each news sample, there are three experimental conditions: presenting only the headline text to the annotator (condition *T*), only the lead image (condition *I*), or the headline and image together (condition *TI*). For each experimental condition, we obtained 10 annotations per sample with each providing: the dominant emotion that the annotator feels among eight emotional categories (Amusement, Awe, Contentment, Excitement, Fear, Sadness, Anger, and Disgust), the intensity of the emotion on a scale of 1–5, and a free-flow English written text describing why the annotator feels that emotion. The overall distributions of responses across the eight emotions in NEMO⁺ are shown in Figure 1.

3.2 Prediction Difficulties in NEMO⁺

We identified some interesting properties in the dataset that make it challenging to predict a single emotional response, which we discuss in detail below. These are intrinsic to the nature of the dataset and are not limitations of the machine learning models benchmarked in this study.

3.2.1 Limited Context Carried in Images

Some of the news images or headlines do not carry much context on their own, like the example shown

in Figure 2. This image provides no clear indication of the identity of the person in the image nor the content of her speech, while the corresponding headline gives more context into the original news content. As we can observe from the viewers’ free-flow responses when presented with only the image (*I* condition), their reported dominant emotions depend largely on speculations. The sample image elicits no negative emotions like sadness, which is present in both the *T* and *TI* conditions. In such news items, the headline text is essential in helping viewers form holistic emotional impressions.

Example 1	
Emma Gonzalez Brought to Tears Honoring Victims of Gun Violence 	
Response Samples	
<i>T</i>	1. "gun violence is such a serious issue" - Sadness 2. "it is sad to know the lost of someone" - Sadness 3. "she is compassionate" - Contentment 4. "I have never been through situation like her but I can imagine how sad it feels" - Awe
<i>I</i>	1. "The person may talks about something about anti gun activity." - Contentment 2. "I think this woman looks full of energy, and she is able to do something to change the situation we face now." - Amusement 3. "this person belongs to the LGBTQ community and is likely to be pro gun controls" - Excitement 4. "I really respect her work as an activist" - Awe
<i>TI</i>	1. "This girl looks so young and honoring victims of gun violence is sad" - Sadness 2. "victims should be remembered" - Sadness 3. "I respect what Emma Gonzalez did and admire her courage to speak for Victims of Gun Violence." - Awe 4. "because the young women in the picture must have spoken about people who were killed by guns in a way that moved the audience deeply, according to the headline" - Awe

Figure 2: News sample among the 1,297 data points in NEMO⁺ with samples of the corresponding emotional responses. The image does not provide enough context of the news.

3.2.2 Emotional Diversity

Another interesting property we observed is that many news items evoke a diverse set of emotional reactions. In the example in Figure 3, annotators have differing emotional reactions towards a given news sample, when presented with the image and headline separately or together. Even positive emotions (Excitement, Awe, Contentment, Amusement) can vary significantly as shown in the example. Moreover, as can be observed from the *T* condition, while written responses suggest annotators agree in a sense, some viewers express negative emotions like anger instead of positive emotions, as they feel that the younger generation should not have to fight for safety.

3.3 Dataset Curation

For the rest of the discussion, let n_{labels} be the number of emotional response types that serve as labels for a news sample and m the number of people that

Example 2	
Teenagers will lead the charge and demand change at anti-gun violence March for Our Lives	
	
Response Samples	
T	1. "it talks of change" - Excitement 2. "teenagers are not supposed to be ones pushing anti-gun-violence, but they are now because the issues are threatening their safety." - Awe 3. "Someone is doing something, but I wish it didn't have to be children" - Contentment 4. "it shouldn't be the students fighting for their safety, it should be their government, parents and schools" - Anger
I	1. "I feel people's emotion toward the gun violence in a passion way." - Amusement 2. "It's good to see the young generation standing up against possession of arms" - Contentment 3. "I respect these people. They try their best to change the world we face." - Awe 4. "young people are speaking up against easy accessibility to gun in the country" - Excitement
TI	1. "teenagers are brave" - Amusement 2. "This campaign is for the good thing" - Excitement 3. "Because This creates awareness about a serious issue." - Contentment 4. "the young women in the image are activists against gun violence." - Excitement

Figure 3: News sample among the 1297 data points in NEmo⁺ with varying emotional response samples in all conditions.

annotate each news sample. We define $\mathbf{v} \in \mathbb{N}^{n_{\text{labels}}}$ to be the frequency annotation vector of a sample, and the entry $v_i \in \{0, \dots, m\}$ describes how many annotators experienced the emotion expressed by the i -th label. To curate the NEmo⁺ dataset for our purposes, we process $n_{\text{labels}} = 8$ possible emotional responses (amusement, awe, contentment, excitement, fear, sadness, anger, disgust); in this order, of $m = 10$ experiment participants. A frequency annotation vector of (0, 0, 1, 0, 0, 2, 0, 7), for example, means that 7 participants experienced the emotion ‘disgust’, 2 the emotion ‘sadness’, and one the emotion ‘contentment’.

In Section 3.2, we observed that the 1,297 news data points of the NEmo⁺ dataset elicited two types of responses: (1) noticeable emotional consensus in the annotations and (2) varying emotional responses with no clear inclination towards a single emotion. We design a subset of the NEmo⁺ dataset, the NEmo⁺-Consensus ("NEmo⁺-C") dataset, that only includes news item with emotional consensus, removing those samples for which people had varying opinions. For this, we experimented with two different filtering methods, discussed below.

3.3.1 Filtering by Rank Diff: Nemo⁺-CR

We defined the rank difference for a news sample to be the difference in frequency between the most frequent emotional response by the group of annotators and the second most frequent emotional response by the group. For the example frequency annotation vector described above, (0, 0, 1, 0, 0, 2, 0, 7), we sort the entries to yield {disgust: 7,

Filter Method	T	I	TI
NEmo ⁺ -CR	365	525	388
NEmo ⁺ -CE	371	514	385
Intersection	199	366	200

Table 1: Filtered data size by filtering method (Rank Difference / Entropy) in all three conditions (T , I , TI). The third column (Intersection) shows the number of samples selected by both filtering methods. We can observe that the I condition is where people have the most emotional consensus in both filtering methods.

sadness: 2, contentment: 1}. Then the rank difference is the frequency difference between the highest ranked emotion ‘disgust’ and the second highest emotion ‘sadness,’ which is 5. This approach is similar to the margin of confidence uncertainty used by Scheffer et al. (2001) as it also examines the difference between the highest and second highest items. In the rank filtering method, we process the NEmo⁺ dataset to only keep news items that have a rank difference of greater than or equal to τ_{rank} . Any news sample, for which the rank difference of the frequency annotation vector lower than τ_{rank} , is removed. We call this filtered dataset Nemo⁺-CR for "Consensus by Rank." We chose $\tau_{\text{rank}} = 3$ to balance having enough consensus in the total $m = 10$ annotations for a particular news sample and having enough data for training machine learning models. The size of Nemo⁺-CR for $\tau_{\text{rank}} = 3$ is shown in Table 1.

3.3.2 Filtering by Entropy: Nemo⁺-CE

The frequency annotation vector can be considered a probability distribution of emotions. If the participants’ emotional responses vary strongly for a news sample, we consider the response uncertain. If there is consensus among the participants, however, we consider the response certain. Since entropy is a measure of the uncertainty of a probability distribution, we can use it to filter the news items. We keep those news items with small entropy values, containing less uncertainty in the emotional distribution of the frequency annotation vector. This is similar in spirit to the rank difference filtering as both methods aim to select those news items that evoke strong emotional consensus. For a fair comparison, we selected the entropy filtering threshold so that the resulting filtered dataset is similar in size to the rank difference-filtered dataset. The size of the resulting filtered dataset Nemo⁺-CE (Consensus by Entropy) is shown in Table 1.

4 Method

We benchmark machine learning models, described in detail in Section 5, for each of the three conditions (T , I , TI) to examine whether text or image when presented separately or together provide more context and help viewers form an emotional response towards particular news content.

4.1 Prediction on the Consensus Data

We performed single label-classification on NEmo⁺-Consensus (Nemo⁺-CR and Nemo⁺-CE). As each news sample has $m = 10$ emotional annotations, we first need to create the single ground truth representative emotion for each news sample. The single representative emotion we use for prediction in the following discussions is simply the most frequent emotion in the m annotations.

4.1.1 Classification on Headline Text

For the T condition, our system aims to predict the single emotional label based on the headline text as the input. This becomes an n_{labels} -class classification task.

4.1.2 Classification on News Image

For the I condition, we developed two separate approaches. The first approach, intuitively, is to predict the emotional label based on the image data itself. However, due to the limited size of the Nemo⁺-CR and Nemo⁺-CE datasets, it is difficult for our system to extract meaningful features from 2-dimensional image data. Furthermore, the images in our dataset do not always provide enough context to the actual content of the news as discussed in Section 3.2.1.



Figure 4: This news image has Web entity tags (concatenated): "Gun Concealed carry Firearm Weapon Gun safety Gun ownership Rifle Semi-automatic firearm Gun control Shooting" and image caption (automatically generated): "A student at the school in Hutsonville, Ill., last week."

In order to infuse some context into image data, we mapped images to text using the Google Web Entity Tagger API ¹ that uses pre-trained models

¹<https://cloud.google.com/vision/docs/detecting-web>

to quickly assign web entity tags and labels to our images (see Figure 4). These tags include textual context of the news that are not always available in the raw images. We also used another image-to-text conversion approach based on the automatic image captioning method by Tourni et al. (2021). After converting the images into textual data, we used the same pipeline as for text classification.

4.1.3 Classification on Image+Text

For condition TI where we are predicting the emotional response of the annotators when presented with both the headline text and the image, we used a multi-modal classification approach where the model learns from both the headline text and the news image to predict the emotional reactions.

4.2 Prediction on the Full NEmo⁺ Data

One limitation of the single-label classification is the reduced dataset from the filtering of the dataset in order to select the dominant "consensus" emotion. The filtering methods mentioned above (rank difference and entropy filtering) aim to select news items that have strong emotional consensus and have a clear dominant emotion. However, most of the time people expressed diverse emotions. In fact, more than 60% of the data in all three conditions in our NEmo⁺ fall into this category of having no clear consensus, as shown by the sizes of the filtered datasets in Table 1 (NEmo⁺ contains 1,297 news items in total). Our approach to the dilemma of having limited consensus in our dataset is multi-label classification. For every news sample, we turned the frequency annotation vector $\mathbf{v}$ from the 10 annotations into a list of binary labels based on a fixed frequency threshold t . We set each entry $v_i \in \{0, \dots, m\}$ of the frequency annotation vector $\mathbf{v}$ to 1 if $v_i \geq t$ and zero otherwise. For example, for a frequency threshold t of 2, we turned the frequency annotation vector $[0, 0, 1, 0, 1, 2, 1, 6]$ into $[0, 0, 0, 0, 0, 1, 0, 1]$.

5 Models

5.1 Text Models

Due to the recent success of Bidirectional Encoder Representations from Transformers (BERT) in the text classification task (González-Carvajal and Garrido-Merchán, 2020), we used BERT (Devlin et al., 2019) for the text classification machine learning models on our emotional consensus dataset. We also experimented with RoBERTa and

observed similar results, so we chose to use the smaller, more efficient BERT model as the main text classification model for news headlines, image tags, and image captions.

Since our dataset is relatively small for training a deep neural network from the ground up, we explored the approach of whether learning from a related domain will be helpful. The ArtEmis dataset provides a foundation for training our baseline models as Achlioptas et al. (2021) used the same eight emotions as Mikels et al. (2005), in addition to a ninth emotion "something else." We removed all records containing the emotion "something else" in the ArtEmis dataset and used the remaining 401,722 data points to train a text-to-emotion baseline BERT model, and then fine tuned it with our consensus data: NEmo⁺-Consensus (Nemo⁺-CR and Nemo⁺-CE). We refer to this model as A-BERT. We also directly fine tuned a BERT-base-uncased² model without pre-training with ArtEmis data for comparison³.

5.2 Image Models

For predicting the emotional response on solely the image data in NEmo⁺-Consensus, we followed the pipeline of the ArtEmis study (Achlioptas et al., 2021) and used a Resnet34 architecture with initial weights that have been pre-trained on the ImageNet dataset with 100,000+ images (Deng et al., 2009) and used the KL-divergence of the frequency annotation vector (from the annotations in the *I* condition) relative to the network output (normalized to a probability distribution) as the loss function.

The output of the model is a distribution of the likelihood of each emotion. We compared the maximum likelihood predicted emotion with the most frequent emotion in the ground truth to measure the performance of the single label prediction. We then fine tuned on the NEmo⁺-Consensus dataset and refer to this model as A-ResNet. We also directly fine tuned an imageNet based Resnet model without pre-training with ArtEmis for comparison.

5.3 Multimodal Image and Text Model

For predicting the emotional response when the viewers are presented with both the headline text

²pre-trained with the weights of the Hugging Face bert-base-uncased model: huggingface.co/bert-base-uncased

³We also experimented with BERT-base-cased model, which is a case sensitive model, and it gave similar results to the uncased model. For the rest of the experiments, we continued using the uncased model.

and the image in NEmo⁺-Consensus (NEmo⁺-CR and NEmo⁺-CE), we fine tuned a BERT based multi-modal bitransformer model introduced by Kiela et al. (2019) using both the headlines and images.

We did not pre-train the multimodal models with ArtEmis because unlike NEmo⁺, there is no single text (i.e., headline) for every image in ArtEmis. Instead, for each image in ArtEmis, there are multiple free flow text responses indicating various emotions. It is not straightforward to choose the "best" text to pair with an image for an indicated emotion as some free flow responses might be better at indicating emotions than others. We leave such exploration for future work.

5.4 Models for NEmo⁺ with Diverse Emotions

Since the NEmo⁺ dataset contains news data points where there is no emotional consensus, we performed multi-label text classification by fine tuning BERT for all three conditions. For condition *I*, we converted the image to textual data using the Google Web Entity Tagger API. For condition *TI*, we concatenated the tagger converted text with the original news headline text as the input to the multi-label model.

6 Evaluation Metrics

6.1 Single-label Classification

The main metric we used for single-label classification is accuracy in predicting the most frequently elicited emotion. Since there are $n_{\text{labels}} = 8$ classes, the expected classification accuracy based on a random guess, i.e., picking a class uniformly at random among all classes (independently for each sample) is given by $1/n_{\text{labels}} = 12.5\%$, a rudimentary baseline for accuracy. However, the NEmo⁺-Consensus dataset is imbalanced towards negative emotions. Therefore, we also compared our models to the majority baselines (the percentage of the dominant emotion in the dataset) to take into account the imbalanced nature of the dataset. These are shown in Table 2. As can be seen in Table 2, rank difference filtering (Nemo⁺-CR) provides a more consistent sample size with emotional consensus across all 3 conditions than entropy filtering.

6.2 Multi-label Classification

For multi-label prediction, we used Hamming distance (Sorower, 2010), exact match accuracy, and

Condition	Nemo ⁺ -CR	Nemo ⁺ -CE
<i>T</i>	41.76%	27.96%
<i>I</i>	41.98%	42.19%
<i>TI</i>	42.27%	37.5%

Table 2: Majority baselines of the Nemo⁺-Consensus dataset (Nemo⁺-CR and Nemo⁺-CE) under each condition. The percentages shown correspond to the fractions of news samples labeled with the dominant emotion in each dataset-condition combination in the test set.

rank-based average precision (LRAP)⁴ to evaluate each model’s predictions.

7 Results

We split the datasets into train / validation / test sets in the ratio of 50%:25%:25% and all of experiment results are reported on the test set.

7.1 Single-label Prediction on Consensus Data

The test time performance of all of our single label prediction models on the data with emotional consensus (Nemo⁺-CR and Nemo⁺-CE) is shown in Table 3. BERT and A-BERT refer to the models with and without pre-training with the Artemis textual data as described in Section 5.1. ResNet and A-ResNet refer to the models with and without pre-training with the Artemis image data described in Section 5.2.

As shown in Table 3, all of the models we benchmarked outperform the majority baselines in Table 2. Our best model (A-BERT on Nemo⁺-CR) surpasses the random baseline significantly by more than 55 percent-points and the majority baseline by 26 percent-points for the *I* condition. When only headlines are used (*T* condition) transfer learning from the ArtEmis textual data improves the accuracy in both consensus datasets. However, when only images are used (*I* condition), transfer learning from the ArtEmis image data improves accuracy only when images in Nemo⁺ are converted to text. This may be due to intrinsic differences between ArtEmis and Nemo⁺ images. Unlike art-centric images of ArtEmis that can intrinsically convey emotional meaning by themselves, images used in news articles may require additional context in the form of web-tagging or image-captioning to leave similar emotional impressions.

We observe that for the single-label prediction task, all the image-only models outperform text-

⁴LRAP: https://scikit-learn.org/stable/modules/model_evaluation.html#label-ranking-average-precision

Dataset:	Nemo ⁺ -CR		Nemo ⁺ -CE	
Model	BERT	A-BERT	BERT	A-BERT
<i>T</i>	56.0%	57.1%	46.2%	51.3%
Model	ResNet	A-ResNet	ResNet	A-ResNet
<i>I</i>	59.7 %	57.4%	63.2%	61.4%
Models	BERT	A-BERT	BERT	A-BERT
<i>I</i> -Tag	64.3%	68.2%	61.7%	60.9%
<i>I</i> -Caption	63.4%	63.4%	54.7%	53.9%
Model	BERT		BERT	
<i>TI</i>	53.6%		40.6%	

Table 3: Classification accuracies of predicting a person’s emotional response on each filtered dataset for all single-label models. The accuracy of the random guessing benchmark is 12.5% and the majority baselines for each condition is shown in Table 2. *I*-Tag and *I*-Caption refer to models where the image data was converted into text using either the Google Web Entity Tagger API or the GVFC’s automatic captioning. All results from this table are from the mode across 30 runs.

only models as well as models for text combined with image in both filtered datasets. Moreover, Table 1 shows that there are more samples with above-threshold consensus for the *I* condition than for the *T* or *TI* conditions. From this, we hypothesize that lead images may be more likely to evoke similar and more-predictable emotional responses in multi-modal gun violence news.

Somewhat surprisingly, the combined text with image *TI* models have the worst performance in both datasets and we discuss possible reasons for this in Section 8.

7.2 Multi-Label Prediction on Nemo⁺

In our multi-label experiment, we controlled for the frequency threshold we used to convert the frequency annotation vectors into binary labels. The higher we set the frequency threshold to be, the easier the task would become, as the converted binary labels would be more sparse and the emotional distribution would be more concentrated.

For multi-label prediction, we are interested in data points with at least two positive binary labels. As shown in Table 4, the percentage of the training data with at least two positive labels decreases as we increase the frequency threshold for binary conversion. We observe that after a threshold of 3, the multi-label learning task becomes insignificant as the training data contains too few qualifying samples. Therefore, we focus on the frequency thresholds of 1, 2, and 3.

We simulated the random baselines by randomly choosing one of the $n_{\text{labels}} = 8$ emotions $m = 10$ times for each of the 1,297 news items and converting the random frequency annotation vector into a

Threshold	T	I	TI
1	99.7%	96.6%	99.4%
2	92.8%	83.1%	89.9%
3	42.0%	33.9%	41.3%
4	5.6%	5.4%	5.5%
5	0.1%	0.4%	0.3%
6	0.0%	0.0%	0.0%
$\vdots$	$\vdots$	$\vdots$	$\vdots$
10	0.0%	0.0%	0.0%

Table 4: Percentage of the data points that contain at least two 1’s after the conversion to binary labels using different frequency thresholds.

list of binary labels given a fixed frequency threshold, as described in Section 4.2. We then compared the random binary labels to the actual binary labels to compute the random baselines’ multi-label performances.

As shown in Table 5, our models consistently outperform the simulated random benchmark in every condition (T , I , TI), at every threshold (1, 2, 3), and under every metric (higher exact match accuracy and LRAP scores and lower Hamming distance loss). Moreover, for every condition and every metric, the absolute performance *improvement* of our models over the random benchmark increases with threshold value and attains the highest improvement at threshold 3.

Thrshld	Rand-Ham	Rand-EM	Rand-LRAP
1	0.47	1.0%	0.59
2	0.46	0.9%	0.49
3	0.39	1.8%	0.42
Thrshld	T -Ham	T -EM	T -LRAP
1	0.35	6.5%	0.81
2	0.26	7.4%	0.72
3	0.17	20.1%	0.67
Thrshld	I -Ham	I -EM	I -LRAP
1	0.35	3.1%	0.78
2	0.26	13.0%	0.71
3	0.15	29.3%	0.69
Thrshld	TI -Ham	TI -EM	TI -LRAP
1	0.35	5.9%	0.8
2	0.27	11.1%	0.7
3	0.16	21.9%	0.64

Table 5: Test-time Hamming distance (Ham) loss (smaller is better), exact match accuracy (EM) (larger is better), and LRAP score (larger is better) of the three conditions T , I , TI with different thresholds for the binary label conversion. The simulated random baselines are called Rand-Ham, Rand-EM, and Rand-LRAP. The results in this table are from a single run as we observed no significant fluctuations among different runs.

At threshold 3, compared to the random baseline, our model’s Hamming distance loss is lower by 0.22, 0.24, and 0.23 points, exact match accuracy is higher by 18, 28, and 20 percent points, and LRAP score is higher by 0.25, 0.27, and 0.22

points, for the T , I , and TI conditions, respectively. In terms of absolute performance, with increasing threshold the Hamming distance and exact match metrics for T , I , and TI improve, but the LRAP metric becomes worse. As the threshold increases there are fewer examples with many labels (see Table 4). A smaller label space makes the classification task “simpler,” but with fewer examples it becomes harder to generalize. Hamming distance and exact match seem to gain more from a reduced label space than they lose due to reduced sample size. The reverse seems to occur for LRAP.

8 Limitations & Future Work

There exist some limitations to our work. Firstly, the multi-modal classification model we benchmark in the TI condition has exhibited lower performance than in the T and I condition (Table 3). This aligns with findings of Wang et al. (2019) that different modalities generalize and fit at different rates and are prone to overfitting due to increased capacity. We also attribute the lower multi-modal prediction performance in the TI condition to the limited size of NEmo⁺-Consensus. It is more difficult for the model to learn enough features from multiple modalities with that amount of data.

One limitation with our multi-label experiment is that the conversion to binary labels causes a loss in relative scale of information among the n_{labels} emotional categories. An alternative approach to this problem in future work could be to model the distribution of emotions for each news sample with a KL-Divergence loss instead.

In future work, we could also derive deeper insights by using the intensity scores we collected in Section 3.1 to predict the strength of emotional responses to news. Another future task is to predict whether a given news headline and/or image will elicit emotional consensus, or result in a divided response among readers. It will also be interesting to study the relationship between emotional responses and the framing of the news and to extend the task to multilingual setting (Akyürek et al., 2020). Finally, we are interested in making the benchmarked systems for predicting emotional responses to news accessible to researchers from a diverse array of disciplines (in similar fashion to the interactive computational framing website: Open-Framing (Bhatia et al., 2021; Guo et al., 2022)) so that researchers from various disciplines can conduct further studies on the potential benefits and

risks of such system.

9 Conclusion

We have shown that we can effectively, to some degree, predict the emotional response to news headline and image using standard text- and vision- classification models. Our work is the novel attempt at benchmarking the task of predicting how exposure to certain textual and visual news affects people’s emotional reactions. This task has wide implications for both news consumers and news professionals. Potential misuses are the possibilities that our tool can be intentionally used to predict and manipulate the emotional reactions of news consumers with specific choices of news headlines and images. However, news editors could aim to avoid sensationalizing their produced media content by using prediction systems like ours. This would be useful in situations where presentation of sensitive news topics (war crimes, terror, etc.) benefits from a more informed selection of image-to-text combinations that can convey important information over sensational, distracting content. Publishers and experts can use this tool to recognize and avoid emotionally-manipulative content. Social media platforms could also use insights on evoked emotion from media in order to predict whether a post is likely to be click-bait. Educators could also use our system for teaching visual media literacy.

10 Ethical Considerations

Our NEmo⁺ is crowdsourced from students through a U.S.-based university in the Northeast. Our dataset may contain certain political and socio-cultural perspective skews given the narrow demographic. As we expand our dataset, we will incorporate annotators from diverse backgrounds while maintaining the annotation quality. We acknowledge that we have received permission to use the BU-NEmo dataset (Reardon et al., 2022), as their data is freely available for the purpose of academic research in our study. Regarding our annotation collection, we ensure we are not knowingly introducing bias to the data nor inflicting any emotional harm on participants or breaching their confidentiality, for which we have obtained IRB exemption approval. We also acknowledge that our use of the ArtEmis dataset is under ArtEmis Terms of Use⁵

⁵https://www.artemisdataset.org/materials/artemis_terms_of_use.txt

that we as researchers use the database only for non-commercial research purposes.

Acknowledgements

This work is supported in part by the U.S. NSF grant 1838193 and DARPA HR001118S0044 (the LwLL program). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of NSF, DARPA, or the U.S. Government.

References

- Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. 2021. [Artemis: Affective language for visual art](#).
- Shakeel Ahmad, Dr. Muhammad Asghar, Fahad Alotaibi, and Sherafzal Khan. 2020. [Classification of poetry text into the emotional states using deep learning technique](#). *IEEE Access*, PP:1–1.
- Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th annual meeting of the association for computational linguistics*.
- Vibhu Bhatia, Vidya Prasad Akavoor, Sejin Paik, Lei Guo, Mona Jalal, Alyssa Smith, David Assefa Tofu, Edward Edberg Halim, Yimeng Sun, Margrit Betke, Prakash Ishwar, and Derry Tanti Wijaya. 2021. [Open-Framing: Open-sourced tool for computational framing analysis of multilingual data](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–250, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. [IEMOCAP: Interactive emotional dyadic motion capture database](#). *Language Resources and Evaluation*, 42:335–359.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. [ImageNet: A large-scale hierarchical image database](#). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Bernd Dudzik, Joost Broekens, Mark Neerinx, and Hayley Hung. 2020. [A blast from the past: Personalizing predictions of video-induced emotions using personal memories as context](#).
- Paul Ekman and W V Friesen. 1971. [Constants across cultures in the face and emotion](#). *Journal of personality and social psychology*, 17 2:124–9.
- Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. [Misinfo reaction frames: Reasoning about readers’ reactions to news headlines](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127, Dublin, Ireland. Association for Computational Linguistics.
- Omar Gambino and Hiram Calvo. 2019. [Predicting emotional reactions to news articles in social networks](#). *Computer Speech & Language*, 58:280–303.
- Santiago González-Carvajal and Eduardo C. Garrido-Merchán. 2020. [Comparing BERT against traditional machine learning text classification](#).
- Lei Guo, Kate Mays, Yiyang Zhang, Derry Wijaya, and Margrit Betke. 2021. What makes gun violence a (less) prominent issue? a computational analysis of compelling arguments and selective agenda setting. *Mass communication and society*, 24(5):651–675.
- Lei Guo, Chao Su, Sejin Paik, Vibhu Bhatia, Vidya Prasad Akavoor, Ge Gao, Margrit Betke, and Derry Wijaya. 2022. Proposing an open-sourced tool for computational framing analysis of multilingual data. *Digital Journalism*, pages 1–22.
- Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. [Target-dependent Twitter sentiment classification](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 151–160, Portland, Oregon, USA. Association for Computational Linguistics.
- Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. [Supervised multimodal bitransformers for classifying images and text](#).
- Weixin Li, Xuan Dong, and Yunhong Wang. 2021. [Human emotion recognition with relational region-level analysis](#). *IEEE Transactions on Affective Computing*, pages 1–1.
- Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. [Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.
- 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*, MM ’10, page 83–92, New York, NY, USA. Association for Computing Machinery.
- Joseph Minkels, Barbara Fredrickson, Gregory Samanez-Larkin, Casey Lindberg, Sam Maglio, and Patricia Reuter-Lorenz. 2005. [Emotional category data on images from the international affective picture system](#). *Behavior research methods*, 37:626–30.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. [MELD: A multimodal multi-party](#)

- dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.
- Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. [Collecting image annotations using Amazon’s Mechanical Turk](#). In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 139–147, Los Angeles. Association for Computational Linguistics.
- Carley Reardon, Sejin Paik, Ge Gao, Meet Parekh, Yanling Zhao, Lei Guo, Margrit Betke, and Derry Tanti Wijaya. 2022. [BU-NEMO: an affective dataset of gun violence news](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2507–2516, Marseille, France. European Language Resources Association.
- Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis, IDA ’01*, page 309–318, Berlin, Heidelberg. Springer-Verlag.
- Mohammad S. Sorower. 2010. A literature survey on algorithms for multi-label learning.
- Isidora Tourni, Lei Guo, Taufiq Husada Daryanto, Fabian Zhafransyah, Edward Edberg Halim, Mona Jalal, Boqi Chen, Sha Lai, Hengchang Hu, Margrit Betke, Prakash Ishwar, and Derry Tanti Wijaya. 2021. [Detecting frames in news headlines and lead images in U.S. gun violence coverage](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4037–4050, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Himil Vasava, Pramegh Uikey, Gaurav Wasnik, and Raksha Sharma. 2022. [Transformer-based architecture for empathy prediction and emotion classification](#). In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 261–264, Dublin, Ireland. Association for Computational Linguistics.
- Jenq-Haur Wang, Ting-Wei Liu, Xiong Luo, and Long Wang. 2018. [An LSTM approach to short text sentiment classification with word embeddings](#). In *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing (ROCLING 2018)*, Hsinchu, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).
- Weiyao Wang, Du Tran, and Matt Feiszli. 2019. [What makes training multi-modal classification networks hard?](#)
- Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. [Building a large scale dataset for image emotion recognition: The fine print and the benchmark](#).
- Haimin Zhang and Min Xu. 2022. [Multiscale emotion representation learning for affective image recognition](#). *IEEE Transactions on Multimedia*, pages 1–1.