Multi-modal Sentiment Analysis

Alejandro Ciuba, Modhumonty Das and **Nick Littlefield**
School of Computing and Information
University of Pittsburgh, Pittsburgh, PA, USA
{alejandrociuba, mod53, ngl18}@pitt.edu

Abstract

The rise of the neural network sparked revolution in the world of machine learning and artificial intelligence. With new techniques being developed frequently, deep neural models have become the standard for many tasks in AI subfields like computer vision, natural language processing and planning. These new models comes at the cost of complexity and data scaling, with newer larger models needing billions of inputs to squeeze out less and less performance. It is from these disadvantages that researchers have began to focus on new ways to embed more information into models: multi-modality. Rather than relying on only text or images, multi-modal models utilize both inputs during decision making. This, in theory, should lead to models with better contextual understanding of a problem. However, this might not always be the case. Our study tests uni-modal and multi-modal modals on a art-based sentiment analysis task on a variety of models to examine how modality changes model performance. We find that the text uni-modal models perform the best and that even non-neural text models outperform state-of-the-art multi-modal models. We hope our work serves as a reminder that adding more modalities does not necessarily correspond to better model performance.

1 Introduction

Sentiment analysis models user sentiments via texts, images and videos. Its applications vary from sociolinguistic research to marketing analytics (Alessia et al., 2015; Nguyen et al., 2016). Sentiment analysis tasks usually focuses on one modality: either text (e.g. food reviews on a website) or image (e.g. pictures of the food in a menu). While this can be considered adequate for many tasks, researchers have started to note the limitations uni-modality brings—namely, the lack of potentially crucial context. For example, a food review might contain an ironic statement that is only apparent with the food review's photos. It is from uni-modality's limitations on context understanding that we see a larger push for multi-modal models, with many believing it is the necessary next step for AI (Bender and Koller, 2020). This creates an interesting research space to examine modality contributions towards model performance. Models can change performance based on modality and how any additional modalities are brought into its representation space.

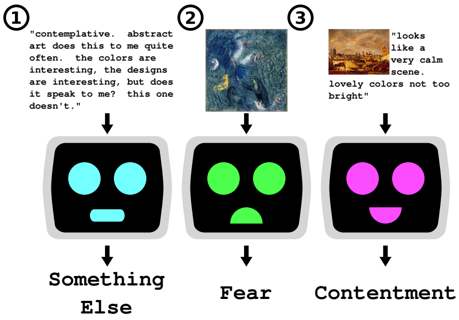
①
"contemplative. abstract art does this to me quite often. the colors are interesting, the designs are interesting, but does it speak to me? this one doesn't."
②
③
"looks like a very calm scene. lovely colors not too bright"
Something Else
Fear
Contentment

Figure 1: An example input and output for the three tested modalities.

1.1 Problem Statement & Research Questions

Our study seeks to understand the differences between modalities and how multi-modality influences model decisions, focusing on sentiment analysis. We ran several smaller-scale uni-modal and multi-modal models which are available on GitHub[1]. Our research questions are:

RQ1 How do uni-modal methods compare when given only text/image input?

RQ2 How do these methods, when adapted to a multi-modal setting, compare with their uni-modal counterparts?

RQ3 How does multi-modality influence our methods?

[1] https://github.com/AlejandroCiuba/CS2756-Group-Project

RQ4 And can we potentially compare our approaches with state-of-the-art multi-modal models?

A brief overview of the modalities tested and their inputs is provided in Figure 1[2][3].

2 Related Work

Existing work emphasizes the importance of multiple modalities for better affective analysis[4] (Shoumy et al., 2020). Achlioptas et al., 2021 created the ArtEmis dataset to develop models for emotion prediction given images and texts. A novel, deep neural network approach to multi-modal sentiment analysis was explored in Hu and Flaxman, 2018. Their goal was to infer latent emotional user states via "emotion word tags" attached to Tumblr posts. The model was validated on an image set used in psychological studies, and the combined multi-modal information was used for the tag prediction task. The effectiveness of multi-modal analysis has also been evaluated by comparing computational emotion classification approaches applied to face videos and bio-sensing modalities in Siddharth et al., 2018, which addressed the limitations of single-sensing modalities, and showed the advantages and increased accuracy of multi-modal affective computing.

3 Dataset

We leveraged the ArtEmis dataset[5] by Achlioptas et al., 2021. ArtEmis is a multi-modal dataset built upon the WikiArt dataset [6]. It consists of over 80,000 paintings spanning 27 styles and 45 genres. Almost 7,000 annotators were asked to evaluate their emotional state (with nine options) upon seeing each painting and then describe why they felt that way. With multiple annotations for most paintings, it totals to 454,684 entries as of April 2024. ArtEmis is a perfect choice for our experiments due to art's subjectivity and its rich annotations.

3.1 Modifications

While ArtEmis is a substantial dataset, there were some issues before we could utilize it in our experiments. First, due to compute limitations, the dataset size needed to be drastically reduced. Secondly, there exists a large class imbalance between the four "positive" (contentment, awe, amusement and excitement) and four "negative" emotions (sadness, fear, disgust and anger) shown in Figure 2. Further complicating matters, there is an inherent imbalance between the amount of text data and image data. The majority of paintings feature more than one annotation (median is five annotations). This means that there is over five times more text information than there is image information per painting. A last major problem, which also arises due to this text-image data imbalance, is the distribution of emotions per painting: some paintings have a high variance in annotated emotions.

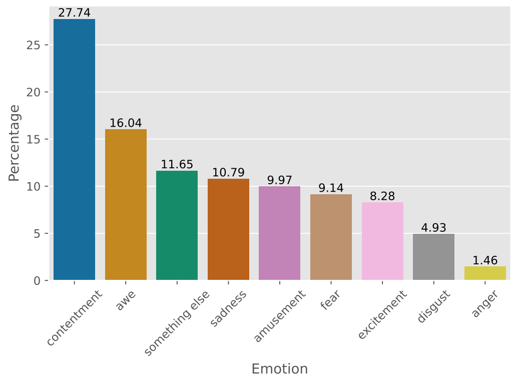

Figure 2: Class proportions prior to modification.

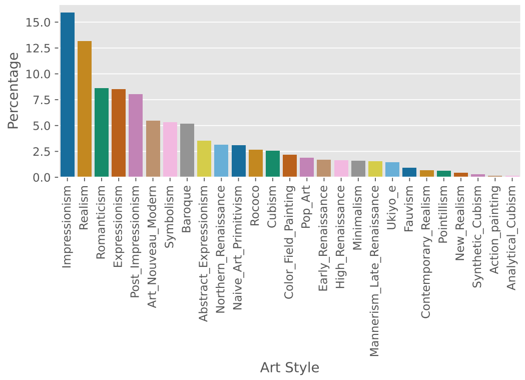

Figure 3: Art style distribution prior to modification

We addressed these concerns in the following order:

1. Randomly sample one annotation per painting, with the weight being proportional to the annotation's token length.

2. Merge the disgust and anger labels to given them more power in our models.

[2] From left to right: "Untitled 1953" by Betty Parsons, "Noah's Ark" by Marc Chagall, "Roofs of Madrid" by Arturo Souto.

[3] All paintings gotten from https://www.wikiart.org.

[4] Sentiment analysis beyond standard positive, negative and neutral tagging (Molina Beltrán et al., 2019).

[5] https://www.artemisdataset.org/

[6] https://www.wikiart.org/

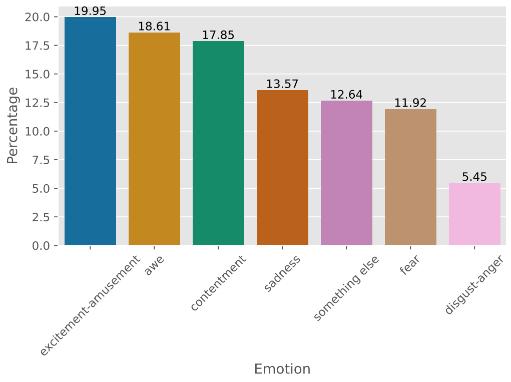

Figure 4: Class proportions after dataset modifications.

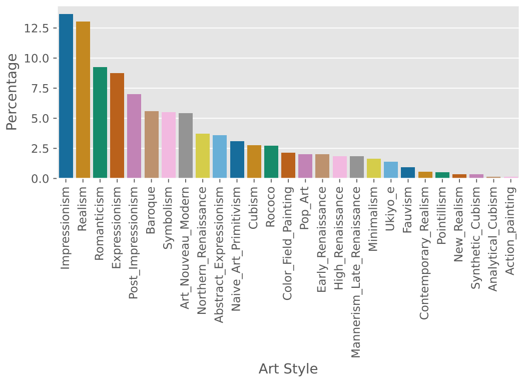

Figure 5: Art style distribution after dataset modification

3. Merge excitement and amusement to keep the positive and negative emotional labels in balance.[7]

4. Pre-downsample the contentment label by a factor of 2.5. This downsizing will not be considered in our models as we are treating this modified dataset as the base.

5. Downsize the dataset to have a 19,000 total examples. Maintain its new emotional label distribution from step four via stratified sampling.

6. Create a 70-20-10 train-test-validation split via stratified sampling.

These splits were used across all models and modalities. Figure 4 shows the new distribution of emotions. This is the same across all dataset splits thanks to the stratified sampling in the final step. Fortunately, other variables, such as art style distribution, did not change much (Figures 3 and 5). We suspect it is because certain art styles were biased towards certain emotional labels, thus, only those biased towards contentment were at risk to be changed.

4 Methodology

Each modality brought its own unique issues and challenges described below. Note that all models, regardless of modality, used the same modified dataset and splits as outlined in 3.1.

4.1 Text Modality

The text modality models utilized only the annotations associated with each painting in a standard 7-way classification task. For the non-neural, non-Bayes models, three versions were tested based on count or TF-IDF vectors of length 5,000 or custom trained doc2vec vectors. The doc2vec model was trained for 70 epochs on the modified training split and outputted a 15-dimensional vector, created with Gensim (Rehurek and Sojka, 2011). We also performed random search hyperparameter fine-tuning for these models. The Bayes family utilized only count and TF-IDF vectors; the feed forward model used either dov2vec or TF-IDF; and the LSTM only worked with word2vec or gloVe vectors (Mikolov et al., 2013; Pennington et al., 2014). All non-neural models were developed using their Scikit Learn (Pedregosa et al., 2011) implementations. The FFNN and LSTM were developed using PyTorch (Paszke et al., 2019). In total we tested 18 text-based modals:

Support Vector Machine (SVM): The SVM tries to find a set of non-linear boundaries between the different categorical groups. The random search sampled from a uniform distribution for its regularization strength and looked sigmoid and radial basis function kernels.

Logistic Regression (LRM): Logistic regression models consist of a linear regressor whose output is fed to a final softmax equation to bound the model's range from $(-\infty, \infty)$ to $[0, 1]$. Our random search sampled the regularization strength and α from a uniform distribution and experimented with lasso, ridge and elasticnet penalties.

Multinomial Naive Bayes (MNB): This offshoot of Naive Bayes is specially adapted to handle n-way classification and has been known to work well with TF-IDF vectors. No hyperparameters were tuned, we used Laplace smoothing.

Compliment Naive Bayes (CNB): An adaption

[7] These were chosen as they were the two smallest positive emotional categories.

to the MNB model by Rennie et al., 2003. Its heuristic allows it to handle class imbalances better while converging more smoothly during training. No hyperparameters were tuned, we used Laplace smoothing.

One vs Rest using CNB (ORC): A one vs. rest model for n-way classification is an ensemble model which takes a provided model and creates n models such that all models compare one label against the "rest". The highest probability is then used for the final output. We went with a MNB model with count vectors after testing.

Feed Forward Neural Network (FNN): Three FFNN variants were created. Two using the doc2vec inputs and one using a 500-length TF-IDF vector. One version of the doc2vec model and the TF-IDF model were trained with oversampling where the probability of sampling entries **i** in batch b was[8]:

$$P(\mathbf{i}) = \left|\left|\frac{1}{|s|^{0.5}}\right|\right| \tag{1}$$

Long Short-Term Memory (LSTM): Our LSTM model was designed to be as similar to the architecture used in the original ArtEmis paper[9]. We reduced the epochs from their original 50 depending on when the loss started to destabilize. Three variants were made: LSTM + GloVe, LSTM + GloVe with oversampling, LSTM + word2vec with oversampling.

DistilBERT: This version of BERT by Sanh et al., 2019 was designed to be lightweight and easy on compute compared to its larger cousins. The model was fine-tuned for 3 epochs using a learning rate of 3e-05 and batch size of 4 (Wolf et al., 2019). It serves as our SOTA baseline for the text modality.

4.2 Image Modality

For the image modality, models utilized only the associated painting for each image. For non-neural models all images of the available paintings were flattened into a one-dimensional vector. All non-neural models were trained using the sklearn learn library. EfficientNet and Vision Transformer were trained using PyTorch.

Logistic Regression Logistic regression is performed on the images by flattening the image into a single vector. Since each image is RGB, it has three channels and each channel is flattened and concatenated together to form this single vector. This vector is then given to make a prediction.

EfficientNet EfficientNet is a family of convolutional neural networks (CNNs) that are designed high accuracy and lower computational cost (Tan and Le, 2020). We utilize EfficientNet-B0 from the huggingface library. This model was fine-tuned in two stages: the first, training only the classification layer, and the second fine-tuning the entire model. We train the model for a total of 10 epochs for the classification layer and 15 for fine-tuning. A batch size of 32 was used with a learning rate of 5e-05.

Vision Transformer (ViT) The vision transformer (ViT) is a transformer architecture designed for computer vision. It takes an image as an input and breaks it down into 16x16 patches and converts these patches into embedding that can be used for classification (Dosovitskiy et al., 2021). We utilize the transformers library to fine-tune a pretrained version of the model. At first all but the classification layer of the model is frozen and trained for 10 epochs. After the model is unfrozen and finetuned for another 10 epochs. Various hyper parameter were tuned for this model. The batch size was 4 and the learning rate used was 5e-05.

4.3 Multi-modality

For the multi-modal modals both the text and images were used to classify a painting as one of the seven available emotions. All models were trained using PyTorch.

Late Fusion Models The late fusion models utilized the fine-tuned EfficientNet and BERT unimodal models trained. For late fusion, the output predictions of EfficientNet and BERT are fused together to make a final prediction (Gallo et al., 2018). The predictions were fused together by taking the average of the two model predictions.

Zero-Shot CLIP Zero-shot CLIP is designed to make accurate predictions without specific training of the target classes (Radford et al., 2021). This approach involves encoding class labels using CLIP's text encoder and then comparing these encodings with the encoded representation of an input image. The comparison used cosine similarity to measure the closeness between the image and each class label. This set of similarity scores is then converted into probabilities using the softmax function, enabling the model to predict the most likely class for the image.

For this study, we explored CLIP's ability to predict classes without prior training, specifically for

[8] PyTorch's `WeightedRandomSampler` handles normalization

[9] https://github.com/optas/artemis/tree/master

the emotion labels from the modified dataset. We encode the names of the classes and compare them through cosine similarity to the images, the class that shows the highest similarity is predicted as the most likely class. The visual encoder used is based on a Vision Transformer (ViT-B/32) architecture [10], which processes images to extract visual features. The textual encoder converts text inputs into embeddings that capture semantic information. The classes are framed as prompts designed to elicit specific emotions, such as "a picture making me feel contentment." Each prompt aims to capture the emotional impact of an image, rather than their specific content.

4.4 Metrics

We analyzed all models across all modalities using accuracy, macro-averaged precision and recall, F_1-score and the macro-averaged false positive rate. For macro-averaged precision, recall and FPR, the formulas for an n-way classification are:

$$\text{Metric} = \frac{\text{Metric}_1 + \ldots + \text{Metric}_n}{n} \tag{2}$$

We wanted both the recall and FPR to get the TPR/FPR ratio, as—due to the imbalanced class distribution—this would be a better representation than the model's accuracy.

In addition, we also computed the weighted precision, recall, F_1-score and FPR for the text modality. The weighted version of these metrics considers the prevalence of each class within the training data (in our case, the whole dataset as the distribution does not change between splits) with the macro-metrics. The formula is:

$$\text{Metric} = \frac{\text{Metric}_1 \times |s_1| + \ldots + \text{Metric}_n \times |s_n|}{n} \tag{3}$$

We used this on the text modality as it seemed to be the best performing modality in our initial testing. The vary between this weighted version and our macro-metrics will help us understand how the class imbalance can affect model performance and if our standard metrics potentially misrepresent modal performance.

5 Evaluation

5.1 Text Modality

The text modality modals fared the best as seen in Table 1. The weighted and macro metrics were similar (Table 2), indicating that our non-neural models are calibrated well to the data distribution. The most surprising result is the fact that the non-neural models significantly outperformed all neural networks besides DistilBERT, which served as our text modality SOTA baseline. Both the FFNN and LSTM performed equivalently to random guessing (Table 3). A lot was tried to get the neural models to improve: adjusting hidden layer sizes, different activation layers (ReLU (Agarap, 2018), Leaky ReLU, hypertangent and softmax), various epoch amounts, batch sizes and learning rates were all adjusted. Sadly, no matter what configuration, the neural models seemed to not fit the data despite a low loss, which we suspect is because the loss does not decrease smoothly. We also suspect that the oversampling did not help because the probabilities between all were very close, meaning a new selection schema could be developed. Work could also be done on optimizing the oversampling scheme. The loss graphs in Figure 6 are also deceiving, since the training accuracy was similar to the testing accuracy.

The LSTM's results were surprising given that a similar model architecture was used in the original ArtEmis paper. There, the authors were able to achieve 63.3% accuracy using text alone. This result was better than the experts, who were only able to identify the correct emotional label 61.2% of the time[11]. Given our study's drastically worse results, we have speculated the following possible issues which could have led to the stark differences. While our LSTM model tries to replicate the majority of their model, there could still be differences which result in the performance. We think this is unlikely, but it is worth mentioning. The biggest factor we believe in the difference of results comes from the two training set sizes and distributions. They deploy an 85-5-10 train-validation-test split. The distribution would have been the same as in Figure 2, with an extreme bias towards contentment, which they do not account for.

5.2 Image Modality

When compared to the text modality, using just the images performs significantly worse. Utilizing just logistic regression performs the same as the neural networks with just the text, with an accuracy of 17%. This isn't surprising as the image is flattened and relations between neighboring pixels

[10] https://huggingface.co/openai/clip-vit-base-patch32

[11] Experts in their analysis were the authors themselves.

Model C | T | D	Accuracy			Precision			Recall			FPR			F1 Score		
SVM	0.56	*0.58*	0.41	*0.61*	0.60	0.43	0.51	*0.55*	0.37	0.08	***0.07***	0.10	0.53	*0.56*	0.36
LOG	*0.59*	*0.59*	0.41	0.61	*0.62*	0.43	*0.57*	0.56	0.37	***0.07***	***0.07***	0.10	*0.58*	0.57	0.37
MNB	*0.60*	0.57	—	0.61	***0.66***	—	*0.57*	0.52	—	***0.07***	***0.07***	—	*0.57*	0.52	—
CNB	***0.61***	0.60	—	*0.60*	*0.60*	—	***0.59***	0.58	—	***0.07***	***0.07***	—	***0.59***	0.58	—
OVR	***0.61***	—	—	*0.62*	—	—	*0.58*	—	—	***0.07***	—	—	*0.58*	—	—

Table 1: Text modality non-neural results. Each result for a model is split as "count | TF-IDF | doc2vec" inputs. **The best results per metric are in bold** while the best result per metric per model are italicized.

Model C | T | D	Precision			Recall			FPR			F1 Score		
SVM	*0.59*	*0.59*	0.42	0.56	*0.58*	0.41	0.10	*0.09*	0.12	0.55	*0.57*	0.40
LOG	*0.60*	*0.60*	0.42	*0.59*	*0.59*	0.41	*0.08*	0.09	0.12	*0.59*	*0.59*	0.40
MNB	0.60	***0.61***	—	*0.60*	0.57	—	*0.08*	0.09	—	*0.59*	0.56	—
CNB	***0.61***	0.60	—	***0.61***	0.60	—	***0.07***	0.08	—	***0.60***	***0.60***	—
OVR	***0.61***	—	—	***0.61***	—	—	*0.08*	—	—	***0.60***	—	—

Table 2: Text modality non-neural weighted results. Each result for a model is split as "count | TF-IDF | doc2vec" inputs. **The best results the metric are in bold** while the best result for a given model are italicized.

Model Macro | Weighted	Accuracy	Precision		Recall		FPR		F1 Score	
FFNN + D2V	0.16	*0.18*	*0.17*	0.14	0.16	*0.14*	0.17	0.13	0.15
FFNN + D2V-O	0.16	0.14	0.15	0.14	0.16	*0.14*	0.18	0.12	0.15
FFNN + TFIDF-O	*0.17*	0.13	0.14	0.14	*0.17*	*0.14*	0.17	0.11	0.13
LSTM + W2V-O	*0.17*	0.15	*0.17*	*0.15*	0.17	*0.14*	0.17	*0.17*	0.19
LSTM + GloVE	0.15	0.14	0.14	0.13	0.15	0.15	0.17	0.14	0.15
LSTM + GloVE-O	*0.17*	0.15	*0.17*	0.14	*0.17*	*0.14*	0.17	0.16	0.18
DistilBERT	**0.67**	**0.68**	—	**0.65**	—	**0.04**	—	**0.65**	—

Table 3: Neural text modality results, **best score for each metric is in bold,** best non-LLM score for each metric is italicized. Accuracy and DistilBERT do not have weighted measurements.

Model	Accuracy	Precision	Recall	FPR	F1 Score
Logistic	0.17	0.15	0.15	0.14	0.15
EfficientNet	0.35	0.30	0.30	**0.09**	0.29
ViT	**0.37**	**0.35**	**0.32**	0.11	**0.32**

Table 4: Image modality results, **best score for each metric is in bold.**

Model	Accuracy	Precision	Recall	F1 Score	FPR
EfficientNet + BERT (Late Fusion)	0.22	0.30	0.22	0.13	0.13
ViT + BERT (Late Fusion)	0.26	**0.42**	**0.25**	0.18	**0.12**
CLIP	**0.28**	0.29	**0.25**	**0.22**	**0.12**

Table 5: Multi-modality results, **best score for each metric in bold.**

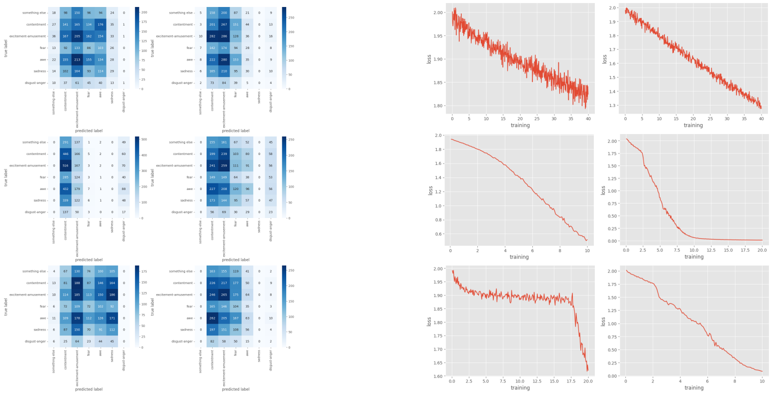

Figure 6: Confusion matrices and losses for all text-based neural models. From left-to-right, top-to-bottom: FFNN + D2V, FFNN + D2V-O, FFNN + TFIDF-O, LSTM + W2V-O, LSTM + GloVE, LSTM + GloVE-O. You can zoom in for a clearer look.

that may help with classification is lost. In total, it confused every class with each other and was basically performing random guessing.

The neural models, however, perform significantly better than the logistic regression model. These models were able to distinguish some of the images for each classes, but still struggled to do it accurately. For EfficientNet, the model was able to correctly classify 35% of the images and the ViT model was able to correctly classify 35% of the images. The differences between these models can be seen clearly from the confusion matrices in Figure 7 The ViT model had the highest precision, recall and F1-Score, but had a higher error rate of 0.11 compared to 0.09 for EfficientNet. The overall results for each of the models are summarized in Table 4.

Given the overall results, however, it is not surprising that the image modality models struggled with predicting the emotion of the painting. This is mainly due to the subjectivity that comes with looking at art work.

5.3 Multi-modality

Late Fusion Models Given the success of the DistillBERT model on accurately predicting the emotion of a painting given the utterance, we expected that combining the output of both the DistillBERT model with the EfficientNet and ViT classifiers could possibly lead to an accurate prediction model for emotion as the output for text combined with the output of the painting would give more information about the overall emotion.

To our surprise, however, the models performed worse than both the unimodal DistillBERT model and the unimodal image models. Using late fusion, the DistillBERT and EfficientNet model had an accuracy of 22% and DistillBERT and ViT had an accuracy of 26%. Surprisingly, however, DistillBERT and ViT do lead to an increase in the precision compared to just the ViT model. Overall, it is unclear as to why these models performed worse when combined, and further investigation is needed. Complete results are show in Table 5.

Zero-Shot CLIP We performed two experiments on Zero-Shot CLIP before and after modifying the dataset with similar test splits used in other modalities. The initial experiment included all nine emotion labels with an accuracy of 36%. Zero-Shot CLIP mostly predicted 'contentment' on a larger scale, and 'fear' and 'something else' on a smaller scale (Figure 8).

However, to make the model consistent with other modalities, a final experiment was performed on the modified dataset, as mentioned in Subsection 3.1. The disgust and anger labels were merged as they were underrepresented, and the excitement and amusement labels were merged to maintain the same number of positive and negative emotion labels. The final experiment's result seemed to be consistent with the initial experiment, where Zero-Shot CLIP was mostly predicting 'contentment' as well as 'fear' and 'something else' in some cases. The only difference noticed was prediction of the merged label 'amusement-excitement' to some extent shown in Figure 9. This experiment led to an overall accuracy of 0.28, as mentioned in Table 5, which is the best among the multi-modal models that we experimented with.

We expected to see better results for our state-of-the-art model, Zero-Shot CLIP, as it takes both visual and textual input into consideration for pre-

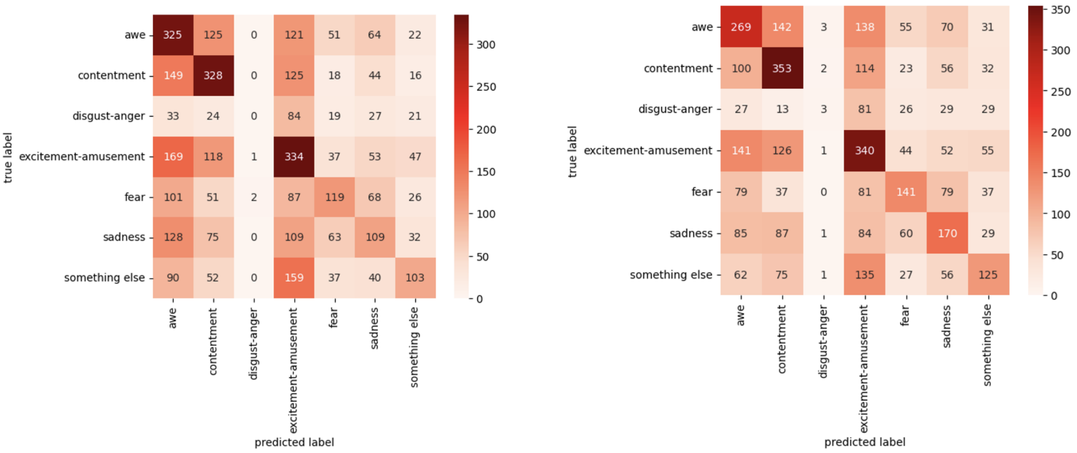

Figure 7: Confusion matrices for the image-based neural models. From left-to-right: EfficientNet, ViT

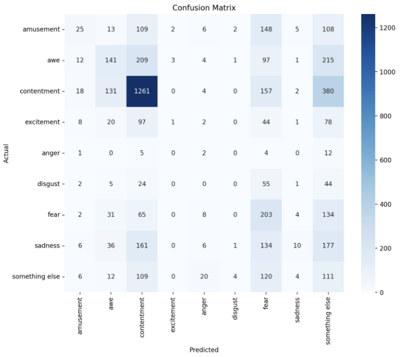

Figure 8: Zero-Shot CLIP Confusion Matrix for Initial Experiment

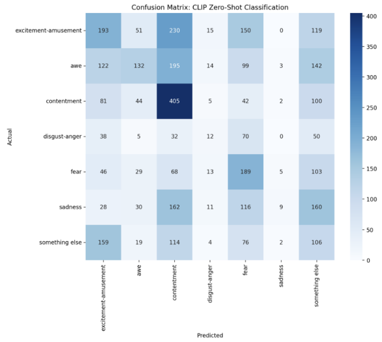

Figure 9: Zero-Shot CLIP Confusion Matrix for Final Experiment

diction tasks. However, it may not be perfect yet for tasks such as emotion or sentiment prediction, as seen in our results. We believe that the arts' subjectivity and lack of information about the emotion labels in textual input are not sufficient to make strong associations between the painting images and the emotion labels. The results are also understandable, as CLIP is highly sensitive to the wording or phrasing of the text input [12]. The word 'something' may be inferred differently in CLIP training, whereas it means an unrecognizable or uncertain emotion in the ArtEmis dataset.

5.4 RQ1: Uni-modal Comparisons

When comparing the non-neural text models against the computer vision models, the text models left the image models in the dust. This result is not surprising considering the text information is directly related to the emotional state of the annotator. The image itself is the art piece, something which most likely had more than one emotion assigned to it by different annotators. While we did account for this in our dataset modification step in Section 3.1, we cannot overcome the simple fact that the art piece itself is going to be inherently more subjective than the annotation. Art can convey a wide range of emotions while the annotations are a concrete description of the annotator's feelings.

5.5 RQ2 & R3: Multi- Vs. Uni-modal Models and Multi-modality Influence

Similar to the text and image modalities, text and multi-modal model difference is still day and night. Text continues to provide the most information out of any modality. We speculate that the image modality actually hinders the capabilities of the multi-modal models given the lack of information the art pieces themselves provide. We believe this is a good example of where and how multi-modality

[12] https://openai.com/research/clip

can fail. We are given concrete evidence of annotator emotion with the annotations, adding in the art itself creates more subjectivity. We hypothesize these models would have performed worse with a training split created from the original dataset as there would then be multiple emotions per painting, further deluding any potential information provided by the paintings; one painting would map to multiple annotations and multiple emotional labels.

5.6 RQ4: The State-of-the-Art

Our state-of-the-art model was Zero-Shot CLIP as described in Section 3.1. It only lost out on precision to the ViT + BERT (Late Fusion) model. But again, it severely underperforms when compared to the text models. Our speculations for this are in the previous Subsection 5.5. CLIP has proven itself to be a SOTA model for many downstream multimodal tasks even with zero and few-shot prompting, but we reiterate our evaluation of the images: much of the arts' features lacks concrete association to certain emotional labels when compared to the text annotations.

6 Limitations & Future Work

There were several limiting factors and challenges we encountered during this study. Lack of computational resources forced us to modify and reduce the dataset's size[13]. It also meant we were unable to fine-tune CLIP, using a modified version of CLIP like Esmaeilpour et al., 2022 or try out more robust, cutting edge models. Challenges setting up HuggingFace training scripts and the image repository caused several delays with the late fusion models, so they remain under-explored. Future work could look at making the models among the modalities more comparable, with more precise and detailed consideration on how the models are fine-tuned across modalities. Work could also be done on examining different multi-modal techniques beyond CLIP and late fusion. More exploratory work could be done analyzing what (if any) painting features are associated with different emotions. Lastly, additional work in the text modality could look at using a different algorithm for choosing the label, given that the probabilities are spread so evenly, such as implementing the algorithm described in Mithal et al., 2017.

[13] Although this led to new and interesting results and a better understanding of proper down-sampling strategies.

7 Conclusion

Our project's goal was to use sentiment analysis to better understand modality's influence in model performance. We ran dozens of models ranging from simple non-neural models, to the SOTA in multi-modality. We found that text-only DistilBERT performed the best across all modalities, but the non-neural text models came close to the results from the orignal paper's experiments. Our version of the LSTM text-only model performed significantly worse then the one from the original paper. We ultimately concluded that the subjectiveness of the art pieces can hinder model performance and that the more concrete, defined information gotten from the text modality leads to this performance gap. We hope our work can be used in the future to inform others on how to approach multi-modality in their model design process.

References

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. 2021. ArtEmis: Affective Language for Visual Art. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11564–11574, Nashville, TN, USA. IEEE.

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*. Citation Key: agarap2018deep.

D Alessia, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. 2015. Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3). Publisher: Citeseer.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. 2022. Zero-shot out-of-distribution detection based on the pre-trained model clip. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(66):6568–6576.

Ignazio Gallo, Alessandro Calefati, Shah Nawaz, and Muhammad Kamran Janjua. 2018. Image and encoded text fusion for multi-modal classification.

Anthony Hu and Seth Flaxman. 2018. Multimodal Sentiment Analysis To Explore the Structure of Emotions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 350–358, London United Kingdom. ACM.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. Citation Key: mikolov2013efficient.

Varun Mithal, Guruprasad Nayak, Ankush Khandelwal, Vipin Kumar, Nikunj C. Oza, and Ramakrishna Nemani. 2017. Rapt: Rare class prediction in absence of true labels. *IEEE Transactions on Knowledge and Data Engineering*, 29(11):2484–2497.

Carlos Molina Beltrán, Alejandra Andrea Segura Navarrete, Christian Vidal-Castro, Clemente Rubio-Manzano, and Claudia Martínez-Araneda. 2019. Improving the affective analysis in texts. *The Electronic Library*, 37(6):984–1006. Publisher: Emerald Publishing Limited.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational Sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830. Citation Key: scikit-learn.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2). Citation Key: rehurek2011gensim.

Jason D M Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. 2003. Tackling the poor assumptions of naive bayes text classifiers. pages 616–623.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. Citation Key: sanh2019distilbert.

Nusrat J. Shoumy, Li-Minn Ang, Kah Phooi Seng, D. M. Motiur Rahaman, and Tanveer Zia. 2020. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *Journal of Network and Computer Applications*, 149:102447.

Siddharth, Tzyy-Ping Jung, and Terrence J. Sejnowski. 2018. Multi-modal Approach for Affective Computing. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 291–294, Honolulu, HI. IEEE.

Mingxing Tan and Quoc V. Le. 2020. Efficientnet: Rethinking model scaling for convolutional neural networks.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Project Contributions

A.1 Alejandro Ciuba

Alejandro worked on all text modality models besides DistilBERT. He also did the initial dataset exploration and produced all graphs seen in the presentation.

A.2 Modhumonty Das

Modhumonty worked on the Zero-Shot CLIP experiments and evaluation for multi-modality.

A.3 Nick Littlefield

Nick worked on DistillBERT for the text modality, all image modality methods, and the late fusion models for multimodality. He also did the preparation of the image dataset for the image modality and the combined multimodality image and text dataset.