

# A Fair and Comprehensive Comparison of Multimodal Tweet Sentiment Analysis Methods

Gullal S. Cheema<sup>1</sup>, Sherzod Hakimov<sup>1</sup>, Eric Müller-Budack<sup>1</sup> and Ralph Ewerth<sup>1,2</sup>

<sup>1</sup>TIB – Leibniz Information Centre for Science and Technology

<sup>2</sup>L3S Research Center, Leibniz University Hannover

Hannover, Germany

{gullal.cheema,sherzod.hakimov,eric.mueller,ralph.ewerth}@tib.eu

## ABSTRACT

Opinion and sentiment analysis is a vital task to characterize subjective information in social media posts. In this paper, we present a comprehensive experimental evaluation and comparison with six state-of-the-art methods, from which we have re-implemented one of them. In addition, we investigate different textual and visual feature embeddings that cover different aspects of the content, as well as the recently introduced multimodal CLIP embeddings. Experimental results are presented for two different publicly available benchmark datasets of tweets and corresponding images. In contrast to the evaluation methodology of previous work, we introduce a reproducible and fair evaluation scheme to make results comparable. Finally, we conduct an error analysis to outline the limitations of the methods and possibilities for the future work.

## KEYWORDS

Multimodal Sentiment Analysis, Information Retrieval, Social Media, Computer Vision, Natural Language Processing, Transformer Models

## 1 INTRODUCTION

Social media has become a phenomenon in terms of its usage by the general public, traditional media, enterprises, and also as a forum for discussing research in academia. With the evolution of the Internet, social media sites, in particular, have become multimodal in nature with content including text, audio, images, and videos to engage different senses of a user. Similarly, sentiment analysis techniques have also progressed from extensively explored text-based [20, 23] to multimodal sentiment analysis [32] of image-text pairs or videos. With two or more modalities, the problem becomes more challenging since every modality might differently influence the overall sentiment, and modalities can have a complex interplay. For image-text pairs, this is even harder as images are perceived as a whole, whereas text is read sequentially. Existing approaches focus on different types of features [24, 36] and complex attention mechanisms [15, 37] to capture the inter-dependencies between image and text to build multimodal models.

Psychological studies have found that human visual attention generally prioritizes emotional content over non-emotional content [5, 9]. A recent study by Fan *et al.* [11] evaluated the inter-relationships of image sentiment and visual saliency in deep convolutional neural network (CNN) models. They proposed a model that prioritizes emotional objects over other objects to predict sentiment, just like human perception. It indicates that to learn a multimodal model for sentiment prediction, visual features should contribute

and consider different objects, facial expressions, and other salient regions in the image. Besides, to learn a multimodal model for sentiment detection, extracted features from two modalities need to be combined in a way that reflects the overall sentiment of the image-text pair. Even though the existing approaches [15, 36, 37] for image-text sentiment detection proposed complex attention mechanisms over different types of features, they fall short on the analysis of features, the number of visual features used, and lack a reproducible evaluation, which hampers the progress in this field.

In this paper, we study the impact of different visual features in combination with contextual text representations for multimodal tweet sentiment classification and present a comprehensive comparison with six state-of-the-art methods. In contrast to previous work, we investigate four different visual feature types: facial expression, object, scene, and affective image content. We utilize a simple and efficient multimodal neural network model (Se-MLNN: Sentiment Multi-Layer Neural Network) that combines several visual features with contextual text features to predict the overall sentiment accurately. In our experiments, we also test the recently proposed CLIP model (contrastive Language-Image Pre-training [27]), which is specifically trained on millions of image-text pairs and reports impressive zero-shot performance on image classification datasets like *ImageNet* [28] and *Places365* [47]. We use this model instead of pre-trained multimodal transformers due to the volume and variety of data it exploited for pre-training, which makes it attractive for different kinds of visual recognition tasks.

We use the publicly available benchmark MVSA [24] (Multi-View Social Data) that consists of two different datasets of tweets and corresponding images. We provide a detailed analysis of both image and text features complemented with an extensive experimental study and outline the limitations of existing approaches. All of the existing approaches for the MVSA datasets use randomly generated, unpublished train and test splits, which makes it impossible to reproduce results or fairly compare them. Thus, we apply k-fold cross-validation so that every sample in datasets is tested once. We share the source code and the new dataset splits used in this paper<sup>1</sup>.

## 2 RELATED WORK

Sentiment detection has been extensively explored for textual social media data, with earlier approaches that were lexicon-based evolving to statistical and machine learning-based classification in the last decade. *SentiStrength* [33] is a well-known lexicon-based approach for short text built using widely occurring words and phrases on social media. Later, Saif *et al.* [29] developed *SentiCircles*

<sup>1</sup>[https://github.com/cleopatra-itn/fair\\_multimodal\\_sentiment](https://github.com/cleopatra-itn/fair_multimodal_sentiment)

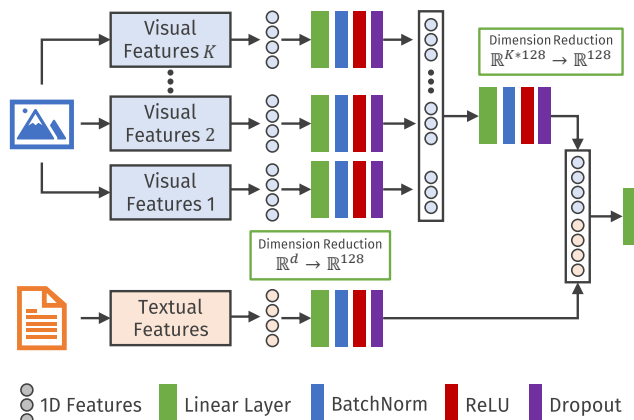
specifically for Twitter sentiment analysis by taking into account the co-occurrence of words in different contexts in tweets. With the prevalence of deep learning, convolutional neural networks (CNNs) [1, 16] and sequential models like Long Short-term Memory (LSTM) [14] networks have been successfully used for tweet sentiment classification. Shin *et al.* [30] developed a hybrid approach by integrating lexicons with a CNN using an attention mechanism. With the rise of image and video data on social media sites like Instagram, Flickr, and Twitter, visual sentiment analysis has attracted a lot of attention recently. Also, in this case, the techniques can be broadly divided into mid-level and deep learning representations. Borth *et al.* [3, 4] developed *SentiBank*, a library to detect visual concepts corresponding to 1200 adjective noun pairs (ANP) for sentiment classification. The concept classifiers rely on low-level visual representations like the “gist” descriptor [25] and wavelet-like Haar features [34]. Yuan *et al.* [43] detected the presence of faces and used facial expressions in combination with low-level visual features to predict the sentiment in images. Using the expressiveness of pre-trained CNNs, You *et al.* [40] fine-tuned the networks for binary sentiment classification on Flickr and Twitter image datasets. Later, they extended the visual sentiment problem to affective image content analysis [41] to predict emotions like *amusement*, *anger*, *awe*, *fear*, *sadness* and *excitement*. Recently, this problem has been explored by several researchers [38, 39, 44, 46, 48] and is discussed in a comprehensive survey by Zhao *et al.* [45].

Multimodal sentiment analysis for social media has gained popularity due to the challenge of combining information from two or more modalities that influence sentiment. Cao *et al.* [7] extracted low-level visual representations, *SentiBank* visual concepts, and lexicon-based features from text to predict sentiment using late fusion strategies. To capture high-level concepts in both image and text, Cai *et al.* [6] as well as Yu *et al.* [42] trained a shallow CNN for text and a deep CNN for images with shared representation to predict sentiment and achieved much better performance than the previous approaches. For the MVSA [24] dataset, in particular, several techniques improved the performance by focusing on cross-modal representations that capture the influence of modalities towards the sentiment. Xu *et al.* [36] proposed an approach using deep CNN with pre-trained features that encode object and scene information from images and aggregated it with contextual *GloVe* [26] (Global Vectors for Word Representation) word embeddings from the text. They used visual feature-guided attention to capture the influence of visual features over word embeddings instead of simply concatenating them for predicting the sentiment. Later, Xu *et al.* [37] proposed a co-memory attention mechanism using similar features to capture the interaction between two modalities and their influence on the sentiment. Recently, Jiang *et al.* [15] proposed another attention mechanism where they used both cross-modal attention fusion followed by modality-specific CNN-gated feature extraction to learn a better representation. They used *ImageNet* [28] pre-trained *ResNet* [13] for visual features, and experimented with *GloVe* [26] and *BERT* [10] (Bidirectional Encoder Representations from Transformers) embeddings for textual features to achieve state-of-the-art results for the MVSA dataset.

### 3 MULTIMODAL SENTIMENT CLASSIFICATION

The main idea is to exploit and investigate different kinds of high-level visual features and combine them with a textual model. The use of channel features as a sequence in conjunction with word embeddings by previous approaches limits their model to two modalities or types of features [15, 36, 37]. In contrast, we aim to investigate the impact of our suggested high-level visual features, which are objects, scenes or places, facial expressions, and the overall affective image content in a more efficient and less complex framework.

A crucial difference between our approach and recent multimodal tweet sentiment approaches [15, 37] is that we use pooling to get one embedding per image instead of using channel features from a pre-trained CNN as a sequence. Similarly, we use a pooling strategy for getting one embedding per a tweet from a textual model. Another difference is that instead of relying on learning bi-attention weights from a limited amount of data, we use a multi-layer neural network to combine different features to influence the sentiment. To investigate the impact of our suggested high-level visual features, we propose a three-layer neural network, where the first two layers aggregate features from different modalities, and the third layer is used for the classification of sentiment. The architecture of our approach is shown in Figure 1. The training details are provided in Section 4. Next, we describe the individual models for each modality and their encoding process to understand the proposed approach.



**Figure 1: Se-MLNN: Proposed architecture for multimodal sentiment classification. Here  $d$  in  $\mathbb{R}^d$  is different for every feature and is provided in Section 3.1, 3.2 and 3.3. Every feature irrespective of the dimension  $d$  is projected down to 128 (First Layer) in order to keep the number of parameters low and not to introduce feature bias. The final layer ( $\mathbb{R}^{256} \rightarrow \mathbb{R}^3$ ) is followed by softmax that outputs the probability of each sentiment.**

#### 3.1 Visual Features

**3.1.1 Object Features ( $E_o$ ):** Different objects in a picture can incite a particular sentiment in a person. For instance, a cute dog or

flowers might bring a positive sentiment, whereas a snake may incite a negative sentiment depending on the context. To encode objects and the overall image content, we extract features from a pre-trained *ResNet* model [13] trained on *ImageNet* [28]. We use *ResNet-50* and its last convolution layer to extract features instead of the object categories (final layer). The final convolutional layer outputs 2048 feature maps each of size  $7 \times 7$ , which is then pooled with a global average to get a 2048-dimensional vector.

**3.1.2 Place and Scene Features ( $E_s$ ):** A scene or a place can also incite different sentiments in a person. For instance, a candy store might bring a positive sentiment, whereas a catacomb might incite a negative sentiment depending on the context. To encode the scene information of an image, we extract features from a pre-trained *ResNet* [13] model trained on *Places365* [47]. In this case, we use *ResNet-101* and follow the same encoding process as above.

**3.1.3 Facial Expressions ( $E_f$ ):** The presence of faces and facial expressions (smiling vs. sad face) in an image can also influence the sentiment in an observer. In the *MVSA* dataset, we found that around 50% of the images contain faces with an average of 2 to 3 faces per image. In order to encode information about facial expressions, we extract the final layer features from a pre-trained [35] *VGG-19* (Visual Geometry Group) model [31] that is trained on around 28 000 [8, 12] face images based on the following seven classes: *angry*, *disgust*, *fear*, *happy*, *sad*, *surprise* and *neutral*. Before extracting the expression features, we first detect faces from an image using a state-of-the-art DSFD [18] (Dual Shot Face Detector) face detector, which are then rescaled to  $48 \times 48$  pixels and input to the *VGG* network. For a given image, if the detector [18] detects  $K$  faces, the *VGG* network outputs  $K$  512-dimensional features that are averaged to get the final feature vector or vector with zeros if no faces are detected.

**3.1.4 Affective Image Content ( $E_a$ ):** Overall affective image content can also be important for multimodal sentiment detection, and research in this area has made rapid progress in recent years with famous datasets from popular social media image sharing websites such as Flickr and Instagram. To encode the overall emotion, we first fine-tune a *ResNet-50 ImageNet* model on publicly available FI (Flickr & Instagram) dataset [41] and extract the last layer convolution features as described above for object and scene embeddings. The dataset consists of around 23 000 training images and eight emotion classes: *amusement*, *anger*, *awe*, *contentment*, *disgust*, *excitement*, *fear* and *sadness*.

## 3.2 Textual Features

Since context and meaning of the words are equally important for the influence of the whole sentence towards the sentiment, we use *RoBERTa-Base* [21] (Robustly optimized BERT approach) to extract contextual word embeddings and employ different pooling strategies to get a single embedding for the tweet. We experimentally found that the average of the last four layers is the most useful and fixed that embedding for all our *RoBERTa* embedding experiments. We finally take an average over the word embeddings to get the single text embedding of 768 dimensions for every tweet. For pre-processing text, we normalize text following [2] and conduct three

experiments by keeping ( $E_T^{+HT}$ ) and removing ( $E_T^{-HT}$ ) the hashtags from the text and also on the raw tweet ( $E_T^{RAW}$ ) text.

## 3.3 Multimodal Features

We use the recently proposed multimodal model *CLIP* [27] that is trained on 400 million image-text pairs collected from the Internet. The model is trained to predict which caption goes with which image and in doing so it learns expressive image representation without the need for millions of labeled training examples. In comparison to multimodal transformers [19, 22, 22], the model uses pairwise learning over  $n$ -pairs of image and text and does not use any cross-attention mechanism to learn multimodal features. This makes the model easy to use as image and text embeddings can be independently computed from the respective image and text encoders. Because of the variety and a large amount of data, the model shows competitive zero-shot recognition performance on 30 different computer vision datasets in comparison to their supervised baselines. This suggests that the amount and quality of visual information encoded in the visual features of the model are much better than the *ImageNet* and *Places365* supervised pre-trained models.

We use a publicly available *CLIP* model<sup>2</sup> and a variant that has visual and textual transformer as image and text encoder backbones. We extract both image and text features from the model where both image ( $C_I$ ) and text embeddings are 512-dimensional vectors. For text, we use the same pre-processing as used for textual models and this results in three types of text embeddings, with hashtags ( $C_T^{+HT}$ ), without hashtags ( $C_T^{-HT}$ ), and raw text ( $C_T^{RAW}$ ).

## 4 EXPERIMENTAL SETUP AND RESULTS

### 4.1 Dataset and Training Details

**4.1.1 Datasets.** We use the *MVSA-Single* (MVSA-S) and *MVSA-Multiple* (MVSA-M) datasets [24] to test our model. The two datasets contain 4869 and 19 598 image-text pairs from Twitter, where both image and text are annotated with a separate label by a single annotator (MVSA-S) or three annotators (MVSA-M with three annotations for each sample), respectively. For a fair comparison and to be consistent with previous work, we process to get the multimodal label and filter the two datasets according to [36, 37], which results in 4511 and 17 025 image-text pairs, respectively. To summarize it, the majority of the assigned class labels over each pair is the pooled label, and the final label falls under three cases: 1) label is valid and same if both have the same label, 2) label is valid and a polar label if either is positive or negative and the other is neutral, and 3) tweet is a conflict and filtered if image and text have opposite polarity labels. The *MVSA-Single* dataset consists of 470 *neutral*, 2683 *positive* and 1358 *negative* samples, while the *MVSA-Multiple* dataset consists of 4408 *neutral*, 11 318 *positive* and 1299 *negative* samples.

**4.1.2 Evaluation and Comparison.** We conduct 10-fold cross-validation where every split ends up with 8:1:1 ratio data and the same label distribution in training, validation and test set. For our ablation study, we report the average accuracy and weighted F1 scores (according to class size) over the 10 folds in Table 1 and 2. For comparison with two other approaches, we report minimum, maximum, and

<sup>2</sup><https://github.com/openai/CLIP>

average accuracy and weighted-F1 measure for all our runs. Our results cannot be directly compared with some previous approaches’ reported results (taken from Jiang *et. al.* [15]) and are only here for reference (marked with †). We evaluate the publicly available *MultiSentNet* [36] implementation<sup>3</sup> and re-implemented the *FENet* [15] model to compare their results (marked with \*) in Table 3.

**4.1.3 Training Details.** We use cross-entropy as an objective function and the Adam (adaptive moment estimation) [17] optimizer for updating the neural network parameters. We observe that the *MVSA-M* dataset label pooling strategy results in noisy labels and to mitigate that we use label smoothing so that the model does not become overconfident. With a smoothing factor of  $\alpha = 0.1$  and  $K = 3$  classes, the one-hot encoded label vector  $y$  becomes:

$$y_{ls} = (1 - \alpha) * y + \alpha / (K - 1) \quad (1)$$

The learning rate is set to  $2 \times 10^{-5}$  and all the models are trained for 100 epochs. We decay the learning rate by a factor of 10 if the validation loss does not decrease for five epochs. A batch size of 32 and 128 is used for *MVSA-S* and *MVSA-M*, respectively. A dropout with the ratio of 0.5 is applied after all the intermediate linear layers to avoid over-fitting. We save the best model (of all epochs) according to the lowest validation loss while training and use it for testing. We use PyTorch<sup>4</sup> for our experiments and extract object features from its publicly available *ImageNet* pre-trained *ResNet-50* [13] model. We train a *ResNet-101* model on the *Places365* dataset and use it for extracting scene features.

**Table 1: Unimodal and multiple visual feature results for *MVSA-Single* and *MVSA-Multiple*. Accuracy and F1 scores are averaged over 10 folds.**

Features	MVSA-Single		MVSA-Multiple	
	ACC	F1	ACC	F1
$E_T^{RAW}$	68.46	66.01	62.83	57.70
$E_T^{+HT}$	68.88	66.49	60.09	55.58
$E_T^{-HT}$	67.50	64.50	65.65	<b>59.61</b>
$C_T^{+HT}$	<b>71.00</b>	<b>68.47</b>	58.52	54.50
$C_T^{-HT}$	68.72	65.59	<b>65.70</b>	59.43
$E_o$	64.69	61.40	65.28	56.63
$E_s$	64.66	61.51	65.13	56.00
$E_a$	64.89	61.65	64.85	56.02
$E_f$	59.63	48.48	<b>66.41</b>	53.21
$C_I$	<b>72.09</b>	<b>70.03</b>	65.42	<b>59.22</b>
$C_I + E_o + E_s + E_f$	<b>70.29</b>	<b>69.51</b>	63.65	59.87
$C_I + E_a + C_f$	69.70	68.66	<b>63.79</b>	<b>60.33</b>
$E_o + E_s + E_a + E_f$	66.42	65.40	63.49	58.58
$E_o + E_a + E_f$	66.10	65.05	63.75	58.89

<sup>3</sup><https://github.com/xunan0812/MultiSentNet>

<sup>4</sup><https://github.com/pytorch/pytorch>

## 4.2 Results

As listing all feature combinations is ineffective, we report the ones which are the most informative and reflect the use of each feature type. Also, we only show a maximum combination with four types of features (visual + textual) as no considerable improvement was observed with five features.

**4.2.1 Unimodal Results:** Table 1 presents the evaluation results of unimodal textual and visual features for the two datasets. For *MVSA-S*, we found out that including hashtag words in the text ( $E_T^{+HT} / C_T^{+HT}$ ) gives slightly better performance than removing hashtags or using raw tweet ( $E_T^{RAW}$ ) text. On the other hand, excluding hashtag words ( $E_T^{-HT} / C_T^{-HT}$ ) from the text works better for the larger *MVSA-M*, where the inclusion of hashtags degrades the average accuracy by almost 6%. The  $E_T^{RAW}$  performance is slightly better than  $E_T^{+HT}$  for *MVSA-M* as *RoBERTa*’s [21] word piece tokenization tokenizes hashtags differently than the pre-processing we used in  $E_T^{+HT}$ . This also shows that for tasks like sentiment detection, pre-processing noisy tweet text can be very crucial. For single visual-only models, we can see that all the visual features except facial expressions ( $E_f$ ) on their own are useful for sentiment detection. For both modalities, *CLIP* features ( $C_I$  and  $C_T$ ) outperform all the other features by 2 to 6% for *MVSA-S* and show similar or slightly better results for *MVSA-M*. This suggests that the pre-training strategy used in *CLIP* learns expressive visual and textual features which can be used in multimodal downstream tasks. Interestingly,  $C_I$  outperforms all other unimodal features for both datasets.

**4.2.2 Visual Combination Results:** Adding any other feature with  $C_I$  degrades the performance indicating that other visual features are not compatible with  $C_I$  in our model, although they slightly increase F1 for *MVSA-M*. With other visual features, we see that the addition of each type of feature (like  $E_o + E_a + E_f$ ) increases the performance in both datasets, especially increase in accuracy and F1 measure by 1% to 4% on *MVSA-S*. The improvement can be attributed to emotion features  $E_a$  and  $E_f$ , which shows that the affective image content is equally important in addition to object and scene information. Facial features on their own perform the worst (very low F1) across datasets as almost 50% of the images have no detectable faces. For *MVSA-M* in particular, the combination of visual modalities only increases the F1 score and needs further analysis.

**4.2.3 Multimodal Results.** Finally, the combination of two modalities increases the sentiment prediction performance across the splits and considerably increases (by 4%) both measures on *MVSA-S* as shown ( $C_I + E_T^{+HT}$ ) in Table 2. This improvement can be seen across all the splits for average, minimum, and maximum values in Table 3, and shows that the addition of modalities not only increases the best score but works for most of the splits. For the *MVSA-M*, the improvement is minimal with 1% accuracy and 3% F1 from unimodal models. Interestingly, visual features combined with *RoBERTa* ( $E_T^{+HT} / E_T^{-HT}$ ) pooled features always outperform combinations with *CLIP*’s text features ( $C_T^{+HT} / C_T^{-HT}$ ) as some are

Table 2: Multimodal results for *MVSA-Single* and *MVSA-Multiple*. Accuracy and F1 scores are averaged over 10 folds.

Features	MVSA-Single		MVSA-Multiple	
	ACC	F1	ACC	F1
	$E_T^{+HT}$		$E_T^{-HT}$	
$E_o$	71.80	70.09	66.28	60.98
$E_s$	72.53	70.77	65.80	60.59
$E_a$	71.98	70.20	66.27	61.13
$E_o + E_f$	72.85	71.57	66.01	62.51
$E_s + E_a$	72.80	71.32	66.12	61.49
$E_o + E_a + E_f$	72.93	71.80	66.31	<b>62.76</b>
$E_s + E_a + E_f$	72.93	71.69	66.19	62.57
$C_I$	<b>75.33</b>	73.76	<b>66.35</b>	61.89
$C_I + E_f$	75.00	<b>73.96</b>	66.08	62.52
$C_I + E_a$	73.95	72.86	65.18	62.03
$C_I + E_s + E_f$	74.73	73.60	66.02	62.51
	$C_T^{+HT}$		$C_T^{-HT}$	
$C_I$	<b>74.97</b>	<b>73.32</b>	<b>66.09</b>	61.27
$C_I + E_f$	74.00	72.58	65.32	<b>61.34</b>
$C_I + E_a$	73.29	72.15	64.68	61.13
$C_I + E_s + E_f$	73.89	72.68	65.43	61.49

shown in two separate grouped blocks in Table 2. This limited performance can be attributed to three issues: 1) considerably higher number of neutral samples that have a higher chance of getting classified as negative or positive, 2) the label pooling strategy [36] used to pool labels from three annotators, which gives preference to positive or negative over the neutral label (refer above) and results in a larger number of disputable labels, and 3) the model’s inability to capture the interactions and differentiate between neutral and polar samples. In the next section, we conduct an error analysis of misclassified samples and group these errors for further consideration. Also, the combination of all four visual modalities in combination with text features did not improve the performance, possibly due to the increase in network parameters.

### 4.3 Error Analysis and Discussion

**4.3.1 Error Analysis.** As stated above, the presence or absence of hashtags has an adverse effect on the performance of both datasets. Further analysis revealed that hashtags that overlap between train, validation, and test splits in *MVSA-S* reflect emotions, and the top-performing splits have a higher overlap of such hashtags. For instance, top overlapping hashtags across different splits in the dataset are words like *love*, *happy*, *passionate*, *winter*, *positive*, *wild*, *strong*, *depressed*, *calm*, *excited*, *broken*, *fear*, *zealous*, *joy*, and *fun*. On the other hand, despite higher overlap between hashtags in *MVSA-M*, the hashtags consist of neutral words like *toronto*, *elxn42*, *yyc*, *cd-npoli*, *nationaldogday*, *job*, *vancouver*, *canada*, *music*, *ottawa*, *hiring*, *realestate*, *realchange*, *halifax*, *photography* and *food*. Although hashtags provide additional context and may reflect sentiment on social

Table 3: Comparison Results for *MVSA-Single* and *MVSA-Multiple*. Results marked with † are taken from Jiang et al. [15], and \* are results of re-implemented models.

Baseline	MVSA-Single		MVSA-Multiple			
	ACC	F1	ACC	F1		
SentiBank & SentiStrength † [4]	52.05	50.08	65.62	55.36		
CNN-Multi † [6]	61.20	58.37	66.30	64.19		
DNN-LR † [42]	61.42	61.03	67.86	66.33		
MultiSentiNet † [36]	69.84	69.63	68.86	68.11		
CoMN(6) † [37]	70.51	70.01	70.57	70.38		
FENet-BERT † [15]	74.21	74.06	71.46	71.21		
Models	ACC			F1		
	Avg	Min	Max	Avg	Min	Max
<b>MVSA-Single</b>						
MultiSentiNet*	63.27	57.87	69.25	59.12	57.83	63.61
FENet-BERT*	69.02	63.76	71.67	67.30	61.42	69.97
Se-MLNN( $C_I + E_T^{+HT}$ )	<b>75.33</b>	<b>70.51</b>	<b>82.04</b>	<b>73.76</b>	<b>69.82</b>	<b>81.14</b>
<b>MVSA-Multiple</b>						
MultiSentiNet*	63.08	54.32	67.10	59.12	54.43	58.57
FENet-BERT*	<b>68.61</b>	<b>61.47</b>	<b>74.40</b>	<b>65.80</b>	<b>60.84</b>	<b>73.56</b>
Se-MLNN( $C_I + E_T^{-HT}$ )	66.35	59.54	70.27	61.89	55.43	65.33

media posts, it is important to understand whether they should be an additional modality to text, filtered, or considered part of the text to better incorporate them in the model. On further inspection of both the datasets, we found around 500 and 700 non-English tweets in *MVSA-S* and *MVSA-M* respectively. Such tweets should either be removed from the datasets or multi-lingual models should be considered for language agnostic sentiment detection. In order to understand the errors from both image and text perspective, we analyze 150 tweets (50 from each class) from the *MVSA-M* that were incorrectly classified by our best model. We divide the errors into three categories.

**Disputable:** We found a considerable amount of disputable labels across all three classes, where some labels are highly disputable and open for discussion. We consider 70 out of 150 labels to be open for discussion, where the majority of them are from positive and negative classes. Quite a few tweets that are disputable in the negative class are selfies of individuals smiling with no negative word mentions in the text. In the positive class, a recurring trait in tweets includes advertisements with no positive connotation in either text or image and thus should be labeled as neutral. For example, both tweet text and image in Figure 2a have neither positive nor negative connotations but it is annotated as *negative*.

**Text in Image:** We also found 66 errors that were due to the embedded text in the image, for which we have not incorporated a visual model in the sentiment prediction. Other common types of images in this category are memes, chat snapshots, weather reports, advertisements, and graphs. For example, text content in the image

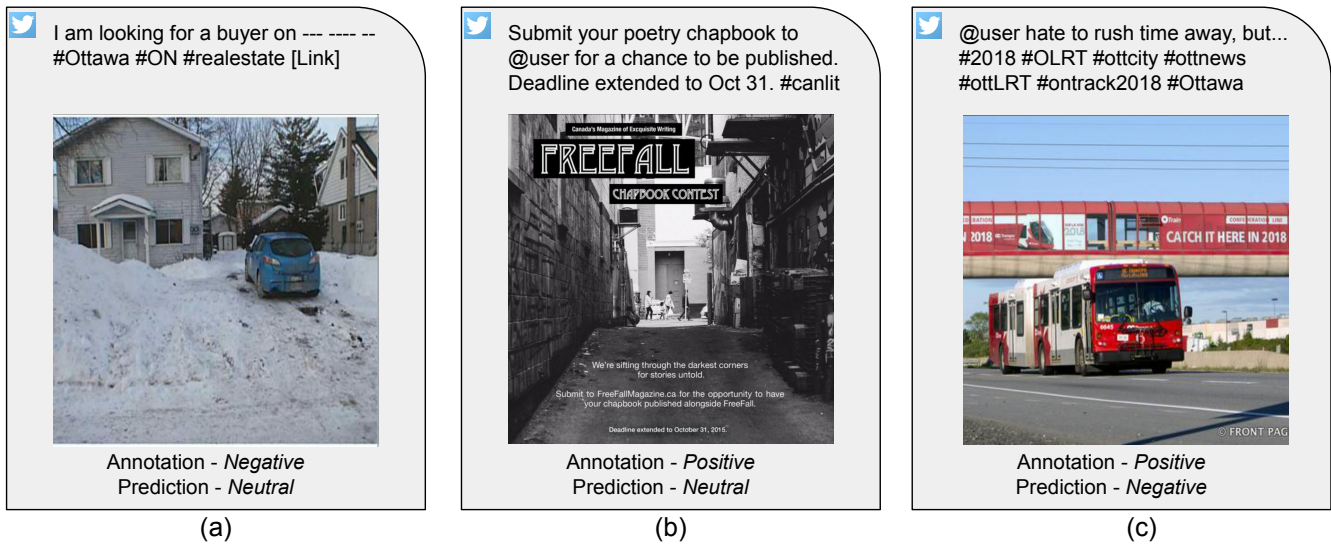


Figure 2: Samples from MVSA-Multiple showing three errors

(see Figure 2b) has important details of a contest that could reflect or change the sentiment of the tweet.

**Missing Context:** Around 67 errors fall into this category where background or cultural information is required to connect the content from both text and image. For example, the tweet text in Figure 2c has a negative connotation, but the image and hashtags refer to city’s bus service among other things that require additional context. Errors, where there is no contextual overlap of image and text or the information is very abstract, are also grouped here.

**4.3.2 Quantitative Examples.** In Figure 3, we show a few examples from MVSA-S where the multimodal model predicts the correct label and unimodal models make an incorrect classification. In the first column (a), where image-only and text-only models predict the correct and incorrect class respectively, texts are mostly neutral and images contain the sentiment specific information. For instance graphic skeletal hand and a cat in the negative and positive example respectively. Some other examples in this category with polar sentiment include graphic images of dead animals and selfies of people with prominent facial expressions. In the neutral category, there are several images of objects (see middle row in Figure 3 a) and advertisements with neutral text. In the second column (b), where image-only and text-only models predict the incorrect and correct class respectively, texts have negative or positive connotation and both modalities need to be seen to predict the correct label. For instance, the negative pair in column (b) has an image of puppies in a cage which does not have a negative connotation, but the text gives important context that the puppies are abandoned. Similar case can be made for the positive pair in the same column. In the last column (c), where both unimodal models predicted the incorrect class, in addition to a few valid cases, there are examples of disputable labels as noted in the previous section 4.3.1. For instance, the positive pair in column (c) has no positive connotation in either text or image and should have a ground truth label as neutral.

## 5 CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a comprehensive experimental evaluation on visual, textual, and multimodal features for sentiment prediction of (multimodal) tweets, including the recently introduced CLIP embeddings. Furthermore, we have compared the performance with six state-of-the-art methods. It turned out that CLIP embeddings can serve as a powerful baseline for the task of multimodal sentiment prediction in tweets. Unlike the used evaluation methodology in previous work, we have introduced a fair and reproducible experimental setup with a 10-fold cross-validation that hopefully provides a useful benchmark for future research and comparison. In future work, we will take cues from the error analysis and focus on models that encode textual content in images as well as contextual information in visual and textual modalities.

## ACKNOWLEDGEMENTS

This work was funded by European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no 812997.

## REFERENCES

- [1] Abdulaziz M. Alayba, Vasile Palade, Matthew England, and Rahat Iqbal. 2018. A Combined CNN and LSTM Model for Arabic Sentiment Analysis. In *Machine Learning and Knowledge Extraction - Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27-30, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 11015)*, Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar R. Weippl (Eds.). Springer, 179–191. [https://doi.org/10.1007/978-3-319-99740-7\\_12](https://doi.org/10.1007/978-3-319-99740-7_12)
- [2] Christos Baziotis, Nikos Pelekis, and Christos Doukeridis. 2017. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch (Eds.). The Association for Computer Linguistics, 747–754. <https://doi.org/10.18653/v1/S17-2126>
- [3] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. 2013. SentiBank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, Alejandro Jaimes, Nicu Sebe, Nozha Boujemaa, Daniel Gatica-Perez,

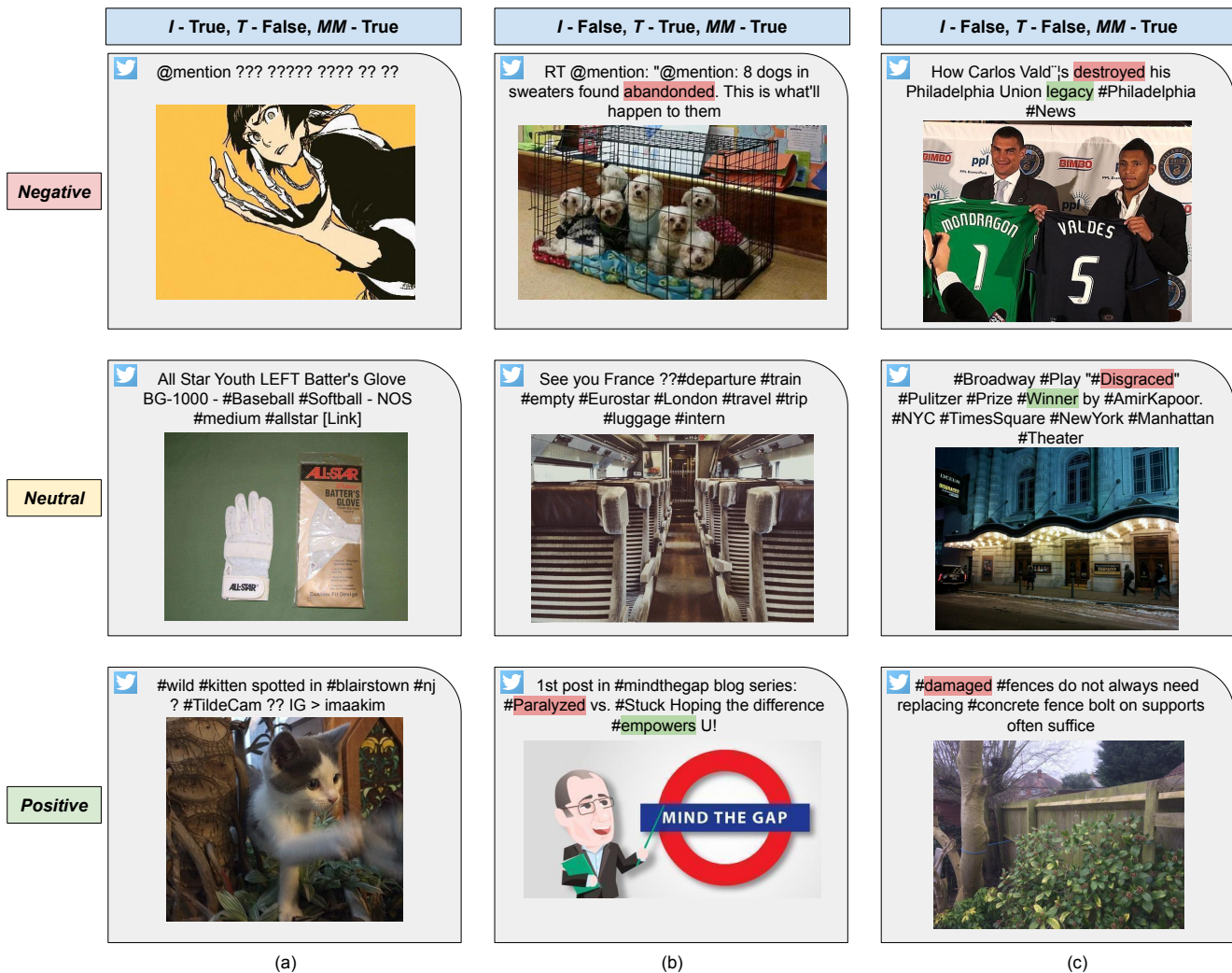


Figure 3: Examples from *MVSA-Single* where multimodal (*MM*) models predict the correct label and either image-only (*I*) or text-only (*T*) models fail

- David A. Shamma, Marcel Worring, and Roger Zimmermann (Eds.). ACM, 459–460. <https://doi.org/10.1145/2502081.2502268>
- [4] Damian Borth, Rongrong Ji, Tao Chen, Thomas M. Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, Alejandro Jaimes, Nicu Sebe, Nozha Boujemaa, Daniel Gatica-Perez, David A. Shamma, Marcel Worring, and Roger Zimmermann (Eds.). ACM, 223–232. <https://doi.org/10.1145/2502081.2502282>
- [5] Tobias Brosch, Gilles Pourtois, and David Sander. 2010. The perception and categorisation of emotional stimuli: A review. *Cognition and emotion* 24, 3 (2010), 377–400.
- [6] Guoyong Cai and Binbin Xia. 2015. Convolutional Neural Networks for Multimedia Sentiment Analysis. In *Natural Language Processing and Chinese Computing - 4th CCF Conference, NLPCC 2015, Nanchang, China, October 9-13, 2015, Proceedings (Lecture Notes in Computer Science, Vol. 9362)*, Juanzi Li, Heng Ji, Dongyan Zhao, and Yansong Feng (Eds.). Springer, 159–167. [https://doi.org/10.1007/978-3-319-25207-0\\_14](https://doi.org/10.1007/978-3-319-25207-0_14)
- [7] Donglin Cao, Rongrong Ji, Dazhen Lin, and Shaozi Li. 2016. A cross-media public sentiment analysis system for microblog. *Multim. Syst.* 22, 4 (2016), 479–486. <https://doi.org/10.1007/s00530-014-0407-8>
- [8] Pierre-Luc Carrier and Aaron Courville. 2013. Challenges in Representation Learning: Facial Expression Recognition Challenge.
- [9] Rebecca J Compton. 2003. The interface between emotion and attention: A review of evidence from psychology and neuroscience. *Behavioral and cognitive neuroscience reviews* 2, 2 (2003), 115–129.
- [10] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. <https://doi.org/10.18653/v1/n19-1423>
- [11] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, and Qi Zhao. 2018. Emotional Attention: A Study of Image Sentiment and Visual Attention. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 7521–7531. <https://doi.org/10.1109/CVPR.2018.00785>
- [12] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron C. Courville, Mehdi Mirza, Benjamin Hamner, William Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Tudor Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. 2015. Challenges in representation learning: A report on three machine learning contests.

- Neural Networks* 64 (2015), 59–63. <https://doi.org/10.1016/j.neunet.2014.09.005>
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016*. IEEE Computer Society, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [14] Minlie Huang, Yujie Cao, and Chao Dong. 2016. Modeling Rich Contexts for Sentiment Classification with LSTM. *CoRR* abs/1605.01478 (2016). [arXiv:1605.01478](http://arxiv.org/abs/1605.01478) <http://arxiv.org/abs/1605.01478>
- [15] Tao Jiang, Jiahai Wang, Zhiyue Liu, and Yingbiao Ling. 2020. Fusion-Extraction Network for Multimodal Sentiment Analysis. In *Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12085)*, Hady W. Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin Pan (Eds.). Springer, 785–797. [https://doi.org/10.1007/978-3-030-47436-2\\_59](https://doi.org/10.1007/978-3-030-47436-2_59)
- [16] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1746–1751. <https://doi.org/10.3115/v1/d14-1181>
- [17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). <http://arxiv.org/abs/1412.6980>
- [18] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. 2019. DSFD: Dual Shot Face Detector. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019*. Computer Vision Foundation / IEEE, 5060–5069. <https://doi.org/10.1109/CVPR.2019.00520>
- [19] Xiujuan Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX (Lecture Notes in Computer Science, Vol. 12375)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 121–137. [https://doi.org/10.1007/978-3-030-58577-8\\_8](https://doi.org/10.1007/978-3-030-58577-8_8)
- [20] Bing Liu and Lei Zhang. 2012. A Survey of Opinion Mining and Sentiment Analysis. In *Mining Text Data*, Charu C. Aggarwal and ChengXiang Zhai (Eds.). Springer, 415–463. [https://doi.org/10.1007/978-1-4614-3223-4\\_13](https://doi.org/10.1007/978-1-4614-3223-4_13)
- [21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). [arXiv:1907.11692](http://arxiv.org/abs/1907.11692) <http://arxiv.org/abs/1907.11692>
- [22] Jiaseen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.), 13–23. <https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html>
- [23] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal* 5, 4 (2014), 1093–1113.
- [24] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmoteleb El-Saddik. 2016. Sentiment Analysis on Multi-View Social Data. In *MultiMedia Modeling - 22nd International Conference, MMM 2016, Miami, FL, USA, January 4–6, 2016, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 9517)*, Qi Tian, Nicu Sebe, Guo-Jun Qi, Benoit Huet, Richang Hong, and Xueliang Liu (Eds.). Springer, 15–27. [https://doi.org/10.1007/978-3-319-27674-8\\_2](https://doi.org/10.1007/978-3-319-27674-8_2)
- [25] Aude Oliva and Antonio Torralba. 2001. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *Int. J. Comput. Vis.* 42, 3 (2001), 145–175. <https://doi.org/10.1023/A:1011139631724>
- [26] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1532–1543. <https://doi.org/10.3115/v1/d14-1162>
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR* abs/2103.00020 (2021). [arXiv:2103.00020](http://arxiv.org/abs/2103.00020) <https://arxiv.org/abs/2103.00020>
- [28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115, 3 (2015), 211–252. <https://doi.org/10.1007/s11263-015-0816-y>
- [29] Hassan Saif, Yulan He, Miriam Fernández, and Harith Alani. 2016. Contextual semantics for sentiment analysis of Twitter. *Inf. Process. Manag.* 52, 1 (2016), 5–19. <https://doi.org/10.1016/j.ipm.2015.01.005>
- [30] Bonggun Shin, Timothy Lee, and Jinho D. Choi. 2017. Lexicon Integrated CNN Models with Attention for Sentiment Analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, Alexandra Balahur, Saif M. Mohammad, and Erik van der Goot (Eds.). Association for Computational Linguistics, 149–158. <https://doi.org/10.18653/v1/w17-5220>
- [31] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). <http://arxiv.org/abs/1409.1556>
- [32] Mohammad Soleymani, David Garcia, Brendan Jou, Björn W. Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image Vis. Comput.* 65 (2017), 3–14. <https://doi.org/10.1016/j.imavis.2017.08.003>
- [33] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment in short strength detection informal text. *J. Assoc. Inf. Sci. Technol.* 61, 12 (2010), 2544–2558. <https://doi.org/10.1002/asi.21416>
- [34] Paul A. Viola and Michael J. Jones. 2001. Rapid Object Detection using a Boosted Cascade of Simple Features. In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), with CD-ROM, 8–14 December 2001, Kauai, HI, USA*. IEEE Computer Society, 511–518. <https://doi.org/10.1109/CVPR.2001.990517>
- [35] WuJie. 2018. Facial Expression Recognition. <https://github.com/WuJie1010/Facial-Expression-Recognition.Pytorch>.
- [36] Nan Xu and Wenji Mao. 2017. MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06–10, 2017*, Ee-Peng Lim, Marianne Winslett, Mark Sanderson, Ada Wai-Chee Fu, Jimeng Sun, J. Shane Culpepper, Eric Lo, Joyce C. Ho, Debora Donato, Rakesh Agrawal, Yu Zheng, Carlos Castillo, Aixin Sun, Vincent S. Tseng, and Chenliang Li (Eds.). ACM, 2399–2402. <https://doi.org/10.1145/3132847.3133142>
- [37] Nan Xu, Wenji Mao, and Guandan Chen. 2018. A Co-Memory Network for Multimodal Sentiment Analysis. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08–12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 929–932. <https://doi.org/10.1145/3209978.3210093>
- [38] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L. Rosin, and Ming-Hsuan Yang. 2018. Weakly Supervised Coupled Networks for Visual Sentiment Analysis. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018*. IEEE Computer Society, 7584–7592. <https://doi.org/10.1109/CVPR.2018.00791>
- [39] Jufeng Yang, Dongyu She, and Ming Sun. 2017. Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19–25, 2017*, Carles Sierra (Ed.). ijcai.org, 3266–3272. <https://doi.org/10.24963/ijcai.2017/456>
- [40] Quanqiang You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25–30, 2015, Austin, Texas, USA*, Blai Bonet and Sven Koenig (Eds.). AAAI Press, 381–388. <http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9556>
- [41] Quanqiang You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12–17, 2016, Phoenix, Arizona, USA*, Dale Schuurmans and Michael P. Wellman (Eds.). AAAI Press, 308–314. <http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12272>
- [42] Yuhai Yu, Hongfei Lin, Jiana Meng, and Zehuan Zhao. 2016. Visual and Textual Sentiment Analysis of a Microblog Using Deep Convolutional Neural Networks. *Algorithms* 9, 2 (2016), 41. <https://doi.org/10.3390/a9020041>
- [43] Jianbo Yuan, Sean Mcdonough, Quanqiang You, and Jiebo Luo. 2013. SentiBite: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM 2013, Chicago, IL, USA, August 11, 2013*, Erik Cambria, Bing Liu, Yongzheng Zhang, and Yunqing Xia (Eds.). ACM, 10:1–10:8. <https://doi.org/10.1145/2502069.2502079>
- [44] Chi Zhan, Dongyu She, Sicheng Zhao, Ming-Ming Cheng, and Jufeng Yang. 2019. Zero-Shot Emotion Recognition via Affective Structural Embedding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 1151–1160. <https://doi.org/10.1109/ICCV.2019.00124>



- [45] Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn W. Schuller, and Kurt Keutzer. 2018. Affective Image Content Analysis: A Comprehensive Survey. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, Jérôme Lang (Ed.). ijcai.org, 5534–5541. <https://doi.org/10.24963/ijcai.2018/780>
- [46] Sicheng Zhao, Zizhou Jia, Hui Chen, Leida Li, Guiguang Ding, and Kurt Keutzer. 2019. PDANet: Polarity-consistent Deep Attention Network for Fine-grained Visual Emotion Regression. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 192–201. <https://doi.org/10.1145/3343031>
- 3351062
- [47] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 6 (2018), 1452–1464. <https://doi.org/10.1109/TPAMI.2017.2723009>
- [48] Xinge Zhu, Liang Li, Weigang Zhang, Tianrong Rao, Min Xu, Qingming Huang, and Dong Xu. 2017. Dependency Exploitation: A Unified CNN-RNN Approach for Visual Emotion Recognition. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, Carles Sierra (Ed.). ijcai.org, 3595–3601. <https://doi.org/10.24963/ijcai.2017/503>