A Fair and Comprehensive Comparison of Multimodal Tweet Sentiment Analysis Methods

Gullal S. Cheema¹, Sherzod Hakimov¹, Eric Müller-Budack¹ and Ralph Ewerth¹,²
¹TIB – Leibniz Information Centre for Science and Technology
²L3S 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 paper¹.

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, Saïf et al. [29] developed SentiCircles

¹https://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 (\( \mathbb{R}^{256} \rightarrow \mathbb{R}^{128} \)) is followed by softmax that outputs the probability of each sentiment.](image)

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 x 7, which is then pooled with a global average to get a 2048-dimensional vector.

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 embedding 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 image 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, 23], 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\(^2\) 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 19598 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 17025 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, 11318 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

\(^2\)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\(^1\) 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 momentum 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:

\[
y_{i} = (1 - \alpha) * y + \alpha / (K - 1)
\]

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\(^4\) for our experiments and extract object features from its publicly available ImageNet pre-trained ResNet-50 [15] model. We train a ResNet-101 model on the Places365 dataset and use it for extracting scene features.

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 RobBERTa’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 \((C_{t} and C_{f})\) outperform all the other features by 2 to 6\% for MVSA-M and show similar or slightly better results for MVSA-S. 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, \(C_{t}\) outperforms all other unimodal features for both datasets.

4.2.2 Visual Combination Results: Adding any other feature with \(C_{t}\) degrades the performance indicating that other visual features are not compatible with \(C_{t}\) 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 \(E_{a} + E_{a} + E_{f}\)) 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 \((C_{t} + E_{t}^{HT})\) 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’s \((E_{t}^{HT}/E_{t}^{HT})\) pooled features always outperform combinations with CLIP’s text features \((C_{t}^{HT}/C_{t}^{HT})\) as some are

---

\(^1\)https://github.com/xuanan0812/MultiSentNet
\(^4\)https://github.com/pytorch/pytorch

---

### Table 1: Unimodal and multiple visual feature results for MVSA-Single and MVSA-Multiple

| Features | MVSA-Single | MVSA-Multiple |
|----------|-------------|---------------|
|          | ACC | F1 | ACC | F1 |
| \(E_{t}^{RAW}\) | 64.68 | 66.01 | 62.83 | 57.70 |
| \(E_{t}^{HT}\) | 68.70 | 64.50 | 65.65 | 59.61 |
| \(C_{t}^{HT}\) | 71.00 | 68.47 | 58.52 | 54.30 |
| \(C_{t}\) | 68.72 | 65.59 | 65.70 | 59.43 |
| \(E_{a}\) | 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 | 66.41 | 53.21 |
| \(C_{t}\) | 72.09 | 70.03 | 65.42 | 59.22 |
| \(C_{t} + E_{a} + E_{a} + E_{f}\) | 70.29 | 69.51 | 63.65 | 59.87 |
| \(C_{t} + E_{a} + C_{f}\) | 69.70 | 68.66 | 63.79 | 60.33 |
| \(E_{a} + E_{a} + E_{a} + E_{f}\) | 66.42 | 65.40 | 63.49 | 58.58 |
| \(E_{a} + E_{a} + E_{f}\) | 66.10 | 65.05 | 63.75 | 58.89 |
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            |
|                    | $E_{HT}^{T}$ | $E_{HT}^{T}$  | $E_{HT}^{T}$ | $E_{HT}^{T}$  |
| $E_o$              | 71.80       | 70.09         | 66.28       | 60.98         |
| $E_s$              | 72.53       | 70.77         | 65.80       | 60.59         |
| $E_d$              | 71.98       | 70.20         | 66.27       | 61.13         |
| $E_o + E_f$        | 72.85       | 71.57         | 66.01       | 62.51         |
| $E_s + E_d$        | 72.80       | 71.32         | 66.12       | 61.49         |
| $E_o + E_s + E_f$  | 72.93       | 71.80         | 66.31       | 62.76         |
| $E_s + E_d + E_f$  | 72.93       | 71.69         | 66.19       | 62.57         |
| $C_I$              | 75.33       | 73.76         | 66.35       | 61.89         |
| $C_I + E_f$        | 75.00       | 73.96         | 66.08       | 62.52         |
| $C_I + E_s$        | 73.95       | 72.86         | 65.18       | 62.03         |
| $C_I + E_o + E_f$  | 74.73       | 73.60         | 66.02       | 62.51         |
| $C_oHT$            | 74.97       | 73.32         | 66.09       | 61.27         |
| $C_I + E_f$        | 74.00       | 72.58         | 65.32       | 61.34         |
| $C_I + E_s$        | 73.29       | 72.15         | 64.68       | 61.13         |
| $C_I + E_o + E_f$  | 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*, *yyz*, *cdnpoli*, *nationaldogday*, *job*, *vancouver*, *canada*, *music*, *ottawa*, *hiring*, *realestate*, *realechange*, *halifax*, *photography* and *food*. Although hashtags provide additional context and may reflect sentiment on social 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
(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) 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

[2] Christos Bzioumis, Nikos Pelekis, and Christos Doullkeridis. 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/S16-2126

[3] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. 2013. SemEval-12 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, Nouta 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.
[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, Benoît 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