ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230503

The Emotions of the Crowd: Learning Image Sentiment from Tweets via Cross-Modal Distillation

Alessio Serra[a,1], Fabio Carrara[b;*,1], Maurizio Tesconi[c] and Fabrizio Falchi[b]

[a]Università di Pisa, Italy
[b]ISTI-CNR, Pisa, Italy
[c]IIT-CNR, Pisa, Italy
ORCiD ID: Fabio Carrara https://orcid.org/0000-0001-5014-5089,
Maurizio Tesconi https://orcid.org/0000-0001-8228-7807, Fabrizio Falchi https://orcid.org/0000-0001-6258-5313

Abstract. Trends and opinion mining in social media increasingly focus on novel interactions involving visual media, like images and short videos, in addition to text. In this work, we tackle the problem of visual sentiment analysis of social media images – specifically, the prediction of image sentiment polarity. While previous work relied on manually labeled training sets, we propose an automated approach for building sentiment polarity classifiers based on a cross-modal distillation paradigm; starting from scraped multimodal (text + images) data, we train a student model on the visual modality based on the outputs of a textual teacher model that analyses the sentiment of the corresponding textual modality. We applied our method to randomly collected images crawled from Twitter over three months and produced, after automatic cleaning, a weakly-labeled dataset of ∼1.5 million images. Despite exploiting noisy labeled samples, our training pipeline produces classifiers showing strong generalization capabilities and outperforming the current state of the art on five manually labeled benchmarks for image sentiment polarity prediction.

1 Introduction

Mining trends and opinions from social networks provides crucial information to help make strategic decisions in various fields. Twitter data, for example, have been used to explain and predict social issues and user opinions on product brands and sales [16, 33], patient reactions to medicines [1], stock market movements [5], political performances and election outcomes [4, 11, 29] and many others. While most research in sentiment analysis from social-network data focused on text, online interactions increasingly involve visual media such as pictures, edited images, and short videos, putting more interest in visual sentiment analysis (VSA). The main issue of state-of-the-art approaches for VSA is their strongly supervised nature: manually labeling images for VSA is costly due to the subjectivity of image interpretation and the viewer's emotional response, thus requiring multiple labelers and limiting the dataset scale to a few thousand samples [19, 44]. Moreover, natural distribution shifts occurring in opinions and trends would require repeating the labeling process periodically, which is unfeasible.

This paper proposes an automated approach to train models for visual sentiment analysis. Specifically, we tackle the problem of predicting the average polarity of sentiments an image evokes to its viewers, usually coarsely estimated as being 'positive', 'neutral' or 'negative'. We propose an approach based on a cross-modal distillation method; a pretrained textual sentiment predictor, acting as the teacher model, is distilled into a visual sentiment predictor using text-image pairs streamed from random-sampled multimodal posts as training samples. The proposed approach is not fully unsupervised but rather based on distant supervision [30], as we assume a pretrained textual teacher model that transfers knowledge to the student visual predictor. However, the availability of self-supervised, easily fine-tunable language models makes it possible to harness the available resources for textual sentiment analysis and transfer their knowledge to the visual domain without additional labeling costs. Moreover, our approach is employable in a continual learning setup, especially if employed with diachronic language models such as TimeLM [26], providing an effective and cheap way to keep sentiment analysis tools up to date.

We apply our approach to random-sampled Twitter posts in three months (Apr-Jun 2022) and show that the obtained visual models outperform the current state of the art in five manually-annotated benchmarks for image sentiment polarity prediction. We also contribute by releasing the code, trained models, and the set collected and preprocessed images (∼1.5M) used in the experimental phase[2].

In summary, we contribute by

- proposing a cross-modal distillation approach to train image sentiment polarity predictors without relying on manually labeled image datasets,
- testing the obtained models on five manually-labeled benchmarks and outperforming the current state of the art in five of them, and
- publicly releasing the code, the trained models, and the collected data (∼3.7M images) used in our experiments.

2 Related Work

Our main focus is purely visual sentiment analysis, where a judgment can be expressed by looking only at image pixels. Other related tasks are also tackled, such as the well-explored textual-based sentiment analysis [24, 2, 34, 20] and directions also exploiting additional

* Corresponding Author. Email: fabio.carrara@isti.cnr.it.
[1] These authors contributed equally.

[2] https://fabiocarrara.github.io/cross-modal-visual-sentiment-analysis

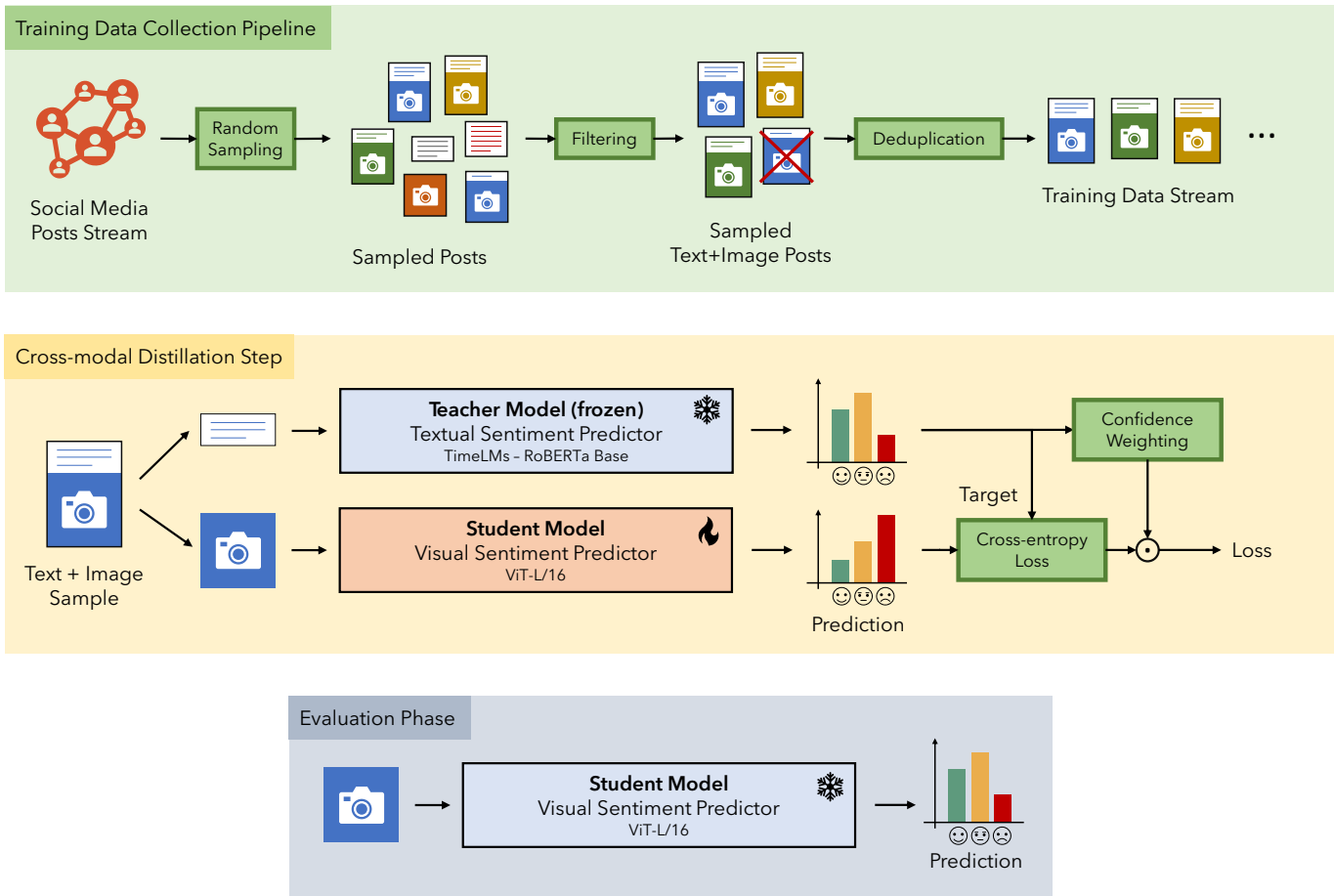

Figure 1. Overview of the proposed approach. In the **Training Data Collection Pipeline**, we use the random 1% of the Twitter global stream as the source of posts. Tweets are filtered and deduplicated to keep text+images samples with long-enough English text, fuelling the **Cross-modal Distillation Step**. Given a text-image pair (i, t), a visual student model (ViT) is trained to predict from the image i the same sentiment polarity of the text t inferred by a textual teacher model (a pretrained textual sentiment classifier, i.e., TimeLMs RoBERTa-Base). Confidence weighting modulates the sample loss with the teacher's prediction confidence. In the **Evaluation Phase**, the frozen student model is applied to images to predict the sentiment polarity using only the visual modality.

inputs or modalities [45, 14, 9, 38, 23, 39] or focusing on aspects different from sentiment, like virality or aesthetics [13, 21, 37].

Visual Sentiment Analysis (VSA) Seminal approaches to visual sentiment analysis, mainly from 2010, were based on extracting handcrafted low-level features from input images based on color, texture, composition, and content characteristics. For example [22] merged SIFT descriptor, Gabor texture, and HSV color histogram to obtain a global feature vector and [27] extracted color, texture and harmonious composition from images. Subsequent approaches leveraged mid-level features of images, such as the one proposed by [6]; they built a visual sentiment ontology consisting of 3'000 adjective-noun pairs that express strong sentiment values and are related to an emotion, represented using the well-known "Plutchik's Wheel of Emotions" psychological model [32]. The adjective-noun pair were used as keywords to get images from Flickr, which were then leveraged to train an individual tracker for each member of the ontology. Subsequently, only reasonably performing detectors were selected to compose SentiBank, their proposed framework capable of extracting mid-level characteristics from images, which can be used as input for a sentiment classifier. However, research has recently been geared towards deep learning models, which can automatically learn how to extract high-level visual characteristics from raw input data. Most methods in this category rely on supervised transfer learning, exploiting various convolutional models like GoogleNet [15], AlexNet-inspired [8], or custom architectures [43]. One of the most recent approaches is [41], which combines global and local features of the image using both a CNN and a saliency detector; in particular, salient sub-images are detected, and then an optimized VGGNet makes a prediction on both the entire image and the sub-images. Finally, the predictions are combined by a weighted sum to detect a positive, neutral, or negative sentiment polarity. Despite having effective models for VSA in the existing literature, their supervised nature usually limits applicability, as preparing training sets is costly, especially when analyzing mutable distributions of data like the one from social network platforms. We tackle this limitation by proposing a cross-modal pipeline capable of training VSA models without requiring human labeling effort.

Dataset for VSA Even for VSA, models are often as good as their training data are. The most used approach to build a dataset for a VSA task relies on manual annotation since it allows getting reliable,

strong labels. However, it is also costly due to the subjectivity of the sentiment we attach to samples, thus requiring more than one annotator to incorporate multiple perspectives into the labeling. For this scope, many researchers [43, 6, 44, 31, 19] relied on crowdsourcing services, i.e., Amazon Mechanical Turk (AMT), to involve multiple labelers and ensure strong labels. In addition, [44, 31] select labelers based on their ability to classify feelings using a qualification test, ensuring cleaner labels. However, scaling datasets beyond the order of tens of thousands of samples still requires a non-negligible effort.

Weak supervision Adopting weak supervision allows us to obtain much larger datasets at the cost of lowering the labeling quality and introducing label noise. In the visual domain, this technique recently gained more and more attention. For example, [36] exploited a complex mixture of raw web signals, connections between web pages, and user feedback to generate a huge image classification dataset, and [28] relied on hashtag prediction on social media images. For VSA, there are just a few examples. The approach of [35] assigns weak sentiment labels to images coming from Flickr based on image tags. Still, it is susceptible to noisy or missing tags and is biased by the tags' choice; similarly, [40] assigns weak labels by analyzing the text content of tweets. Our approach follows this direction by crawling randomly sampled multimodal data from social media streams, but the supervision signal is obtained by distilling a textual sentiment predictor into a visual model.

3 Methodology

As done in previous work, we formulate image sentiment polarity prediction as an N-way image classification problem. Our objective is to learn an image classifier that assigns the correct sentiment label out of N possible labels to an input image without resorting to supervised training and, thus, an expensive manual annotation of images. To do so, we propose an automatic approach organized in two steps; a) *data collection, filtering, and deduplication* and b) *cross-modal distillation*. Figure 1 schematizes our proposal. We further describe each step in the following subsections.

3.1 Data Collection, Filtering, and Deduplication

This first step aims to construct a data stream to fuel the subsequent learning step. We crawl data from a social network of interest by collecting random posts in a specified period. In this work, we demonstrate our proposal on Twitter, but in principle, any platform providing access (free or paid) to large volumes of randomly-sampled posts can be used.

To subsequently apply a cross-modal paradigm, we are interested in filtering out samples having only a single modality in favor of ones containing both text and one or more images. We apply the same filtering steps applied in [40] and keep only tweets that a) have a text comprised of 5 or more words in the English language, b) have at least one image, and c) are not retweets. We thus obtain a set $S = \{s_j\}_{j=1}^{M}$ of text-image pairs $s_j = (t_j, i_j)\,, t_j \in T\,, i_j \in I$, where T and I respectively indicate the space of texts and images. We indicate with M the number of samples at the end of the collection campaign, but in an online learning configuration, S constitutes an infinite data stream.

Due to the virality of some contents, a non-negligible part of posts and corresponding images crawled end up duplicates or near-duplicate images. To make the process leaner and obtain a more varied stream of visual data, we drop samples having the same or nearly-same content in the visual medium. Specifically, we assume two samples $s_1 = (t_1, i_1)$ and $s_2 = (t_2, i_2)$ are duplicates if $\cos(\Phi(i_1), \Phi(i_2)) > \tau$, where $\Phi(i) \in \mathbb{R}^n$ is a feature vector extracted from the image i by a general-purpose pretrained visual model Φ, and τ is an empirically-chosen threshold.

3.2 Cross-modal Distillation

We set up a cross-modal student-teacher learning paradigm fed by data streaming from the previous step.

Let $g : T \rightarrow [0, 1]^N$ a pretrained textual sentiment polarity predictor that maps an input text into an N-dimensional categorical distribution and similarly, $f : I \rightarrow [0, 1]^N$ an image classifier sharing the same label space as g. Given a set of multimodal samples $S = \{s_j\}_{j=1}^{M}$, we train the student model f to align its prediction on the visual modality to the ones of the teacher model g on the textual modality. To do so, we define a confidence-weighted cross-entropy distillation loss as follows. Formally, for a single text-image pair $s = (t, i)$, we minimize the following loss

$$\mathcal{L}(t, i) = \lambda(g(t)) \sum_{k=1}^{N} g_k(t) \log(f_k(i))\,, \tag{1}$$

where $g_k(t)$ and $f_k(i)$ indicate the k-th output of the teacher and student model, respectively, and

$$\lambda(g(t)) = \begin{cases} 1 & \text{if } g_{\bar{k}}(t) \geq c_{\bar{k}}\,,\ \bar{k} = \underset{k}{\operatorname{argmax}}\ g_k(t) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

is a multiplier that weights the sample contribution based on the teacher's confidence. Specifically, the formulation in Equation 2 represents a hard-gating strategy that filters out low-confidence samples, as it sets the sample loss to zero if the probability of the most confident class $g_{\bar{k}}(t)$ is below a predefined threshold $c_{\bar{k}}$ that is defined for each possible class $\{c_j\}_{j=1}^{N}\,, c_j \in [0, 1]$. We select a hard-gating strategy for its simplicity and reasonable effectiveness in our experiments. However, we define $\lambda(g(t))$ to represent a generic weighting scheme for training samples; in future work, we plan to explore other formulations, such as soft gating. During training, the teacher model g is frozen, and only f is updated by gradient-based optimization until convergence.

4 Experiments

4.1 Experimental Setup

Data Collection We collected roughly 3M tweets with 3.7M images (1.26 images per tweet on average) in three months between April and June 2022. Crawling was implemented via the Twitter API Volume Streams[3] that provides a streaming endpoint delivering roughly a 1% random sample of the global and publicly available tweets in real-time. For deduplication, we choose an ImageNet-pretrained ResNet-50 as feature vector extractor Φ; specifically, we use the max-pooled output of the sixth residual block as feature vector and mark tweets as duplicates if their image contents have very-high cosine similarity ($\tau = 0.98875$). Deduplication yielded a $\sim$22% reduction of the image set, which went from $\sim$3.7M to $\sim$2.9M.

[3] https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/introduction

In Table 1, we report a summary of collected data broken down by the three sentiment polarity classes induced by the teacher model chosen in our experimentation (more on this in the following subsections). As an additional source of samples, we also employ B-T4SA [40] — a set of 470586 text+images tweets collected following the same crawling rules between July and December 2016. Following a chronological order, we refer to the B-T4SA dataset as A and our newly collected dataset as B.

Teacher Architecture Among many approaches proposed in the literature for textual sentiment analysis, for the teacher model, we choose a model from Time-LMs [26] — a family of models trained with a *continual learning* approach. It comprises a BERT-based model trained on real-time Twitter data and periodically released, enabling diachronic specialization that is particularly relevant in the social media domain where the topic of discussion changes rapidly, as well as slang and language used. For instance, a model trained before 2019 would not be aware of the meaning of neologisms such as *"COVID-19"* or the different feelings related to *"swabs"* or *"variant"* that we give after the pandemic.We select the Time-LM model released at the end of June 2022, fine-tuned for sentiment analysis on the TweetEval benchmark [3] available in the TweetNLP library [7] This choice also sets the granularity of the prediction ($N = 3$), as the model has three possible outputs; 'positive', 'neutral', or 'negative' sentiment polarity.

Student Architecture As the visual student model, we select a Vision Transformer (ViT) [12] with the final head adjusted to output $N = 3$ logits. We start training from the publicly available checkpoints pretrained on *Imagenet-21k* and on *Imagenet-1k*. During training, we employ data augmentation on the visual pipeline by applying random horizontal flips, shifts, and rotations. Optimization is carried out using the RAdam optimizer [25] with an initial learning rate of 10^{-4}.

Table 1. Summary on collected and preprocessed training data broken down by the sentiment polarity assigned by the teacher model (TimeLM Jun-2022 fine-tuned on TweetEval).

	Collected		Deduplicated
Sentiment	# tweets	# images	# images
Positive	1 206 158	1 593 484	1 299 916
Neutral	1 403 683	1 708 195	1 293 259
Negative	356 002	433 172	329 395
Total	2 965 843	3 734 849	2 922 568

4.2 Benchmarks

To test the effectiveness of the proposed cross-modal training process, we evaluate our models on the benchmarks for image sentiment polarity prediction manually annotated via Amazon Mechanical Turk (AMT). We consider a) Twitter Dataset (TD) [43], b) Flickr&Instagram (FI) [44], and c) EmotionROI [31]. TD provides three benchmarks corresponding to three different levels of label agreement, i.e., where at least five, four, or three AMT workers agreed on the labels assigned to images. The other datasets provide a single set of images with already aggregated labels. TD provides binary labels ('positive' or 'negative') for sentiment polarity. Thus we mask the neutral class output of our models and take the maximum confidence among positive and negative outputs.

Table 2. Summary of manually annotated benchmarks used for evaluation. We report the number of original classes (Classes), the number of Amazon Mechanical Turk involved in labeling (Raters), and the number of images. We report the number of samples remapped into 'positive' and 'negative' polarity for datasets with more than two classes.

			# Images		
Dataset	**Classes**	**Raters**	🙂	🙁	Tot.
Twitter Dataset [43]	2	5	769	500	1,269
EmotionROI [31]	6	432	660	1,320	1,980
Flickr&Instagram [44]	8	1,000	16,430	6,878	23,308

FI and EmotionRoI provide fine-grained sentiment annotations and are used in literature as sentiment polarity benchmarks by mapping labels into two 'positive' and 'negative' polarities [41]. In particular, for dataset FI, the emotions of *Awe*, *Amusement*, *Excitement*, and *Contentment* are mapped to the 'positive' polarity while *Fear*, *Disgust*, *Sadness*, and *Anger* to 'negative'. For EmotionROI, *Anger*, *Disgust*, *Fear*, and *Sadness* are relabeled as 'negative', and *Joy* and *Surprise* as 'positive'. Table 2 reports the characteristics of each dataset, and Figure 2 shows some examples. We adopt TD for preliminary experiments and ablation studies while we compare the best-performing models with other state-of-the-art methods on all the mentioned benchmarks.

4.3 Ablation study

In this section, we evaluate how aspects such as data freshness, data filtering, and model architecture can affect the effectiveness of trained models. We perform experiments varying the inputs and hyperparameters of our approach and producing several models. We apply the obtained models on the TD benchmark in a zero-shot configuration (no learning on benchmark data is performed) and measure the classification accuracy. In Table 3, we report each tested configuration (labeled from 3.1 to 3.8) together with the obtained results we discuss below.

Confidence Filtering In experiments 3.1 and 3.2, we fix the input set (A) and the student model architecture (ViT Base with 86M parameters and a patch size of 32) and run our pipeline with or without confidence filtering, i.e., setting $c_j = .70\,, \forall j$ or $c_j = 0\,, \forall j$ in Equation 2. Comparing rows 3.1 and 3.2 in Table 3, we note that masking low-confidence samples in the student loss helps increase accuracy by 1–2%.

Input Data In experiment 3.3, we repeat experiment 3.2, swapping the set A collected in 2016 with the one collected by us in 2022 (B). We observed a small accuracy loss in the five-agree benchmark. Despite having more images, our set is more unbalanced towards positive and neutral classes with respect to A, which the original authors already balanced during data cleaning. Indeed, setting higher confidence thresholds for those classes (experiment 3.4) mitigates this problem and provides additional improvements also to the lower-agree benchmarks. Combining the two sets (experiment 3.5) further increases performance by ∼2%.

Student Architecture We evaluate scaling the model parameters and the patch size of the student ViT architecture. Starting with the

Figure 2. Samples from the manually-annotated benchmark used for evaluation. From left to right, we show a positive and a negative sample for TD, EmotionROI, and FI benchmarks.

Table 3. Ablation study. Accuracy of sentiment prediction on the three Twitter Dataset benchmarks (at-least-five-, four-, and three-agreement subsets). Experiments 3.1-3.2 show the effects of confidence filtering. Confidence filtering columns report the values used for the parameters $\{c_j\}_{j=1}^{3}$ in Equation 2. Experiments 3.4-3.5 investigate training data collection period and scale. A = Set of tweets collected in Jul-Dec 2016 by [40]. B = Set of tweets collected in Apr-Jun 2022 by us. Experiemnts 3.6-3.8 show the effect of model size and input patch size. The model column indicates the student architecture, i.e., a Base or Large ViT with input patch size of 32 or 16.

		Confidence Filter				Twitter Dataset		
#	Dataset	🙂	😐	☹	Student Model	5 agree	≥4 agree	≥3 agree
3.1	A	-	-	-	B/32	82.2	78.0	75.5
3.2	A	.70	.70	.70	B/32	84.7	79.7	76.6
3.3	B	.70	.70	.70	B/32	82.3	78.7	75.3
3.4	B	.90	.90	.70	B/32	84.4	80.3	77.1
3.5	A+B	.90	.90	.70	B/32	86.5	82.6	78.9
3.6	A+B	.90	.90	.70	L/32	85.0	82.4	79.4
3.7	A+B	.90	.90	.70	B/16	87.0	83.1	79.4
3.8	A+B	.90	.90	.70	L/16	87.8	84.8	81.9

configuration of experiment 3.5, in experiment 3.6, we swap the student model for the larger ViT-Large (307M parameters, ∼3.5x more than ViT-Base), while in experiments 3.7 and 3.8, we repeat experiments 3.5 and 3.6 decreasing the input patch size from 32 to 16 (4x larger input sequences). Decreasing patch size alone (3.7) is more effective than increasing model parameters (3.6), as the visual model can grasp finer details of the input image. Scaling both dimensions together (3.8) produces our best-performing configuration, confirming recent findings [18].

4.4 Comparison with State of the Art

We compare our best model (ViT-L/16 trained on A+B) to state-of-the-art methods on the five manually-labeled benchmarks for image sentiment polarity described in Section 4.2. For a fair comparison, we follow the evaluation protocol of previous work [41] that includes fine-tuning the models on the benchmark data. Specifically, for TD and Emotion ROI, 5-fold cross-validation is performed, while for FI, models are trained on five random splits with 80/5/15 proportions of training/validation/test subsets. For each benchmark, we measure the mean and standard deviation of the accuracy on the test splits.

As seen in Table 4, our models outperform or are comparable to other state-of-the-art methods in all benchmarks. Without fine-tuning, our models still obtain satisfactory results. For the TD benchmark, which shares a data distribution similar to the one of the crawled data used, our model achieves an accuracy comparable to fine-tuned state-of-the-art models, even outperforming them on the 3-agreement subset. On the other hand, the distribution shift between Twitter images and the Emotion ROI and FI benchmarks are too significant to ensure generalization. We deem the culprit to be the class distribution for Emotion ROI, which privileges a negative sentiment polarity contrarily to other datasets, and the domain gap for FI, where images comprise more high-quality artistic pictures rather than synthetic/edited images and pictures taken with a smartphone. However, fine-tuning reduces these gaps, showing that the knowledge in our model can be easily transferred to other domains.

In Figure 3, we report some cherry-picked failure cases of our best model (non-finetuned ViT-L/16) on the Twitter Dataset benchmark. Most failure cases comprise very subjective samples, for which the correct label is not immediately clear, even for a human judge.

5 Conclusion

We presented an automated approach to obtain trained models for visual sentiment analysis targeted for social media mining. Harnessing existing resources for textual sentiment analysis, the proposed cross-modal distillation approach can produce robust models for image sentiment polarity prediction without any human intervention in data collection or labeling. The presented pipeline enables the production of visual diachronic models that capture new concepts and concept drift; our pipeline can be run off-line periodically or on-line via continual learning from streaming social media data in an autonomous way. The experimental phase on Twitter data showed that our models reached a significant performance on manually-annotated benchmarks, setting the new state of the art on five of them. All the collected data, the annotated datasets, and the trained models will be publicly available.

However, several limitations remain to be tackled. One of the main issues (and thus motivation for future work) is the subpar zero-shot generalization to other domains, i.e., social media. Although finetuning our models demonstrated a great transferability of the knowledge

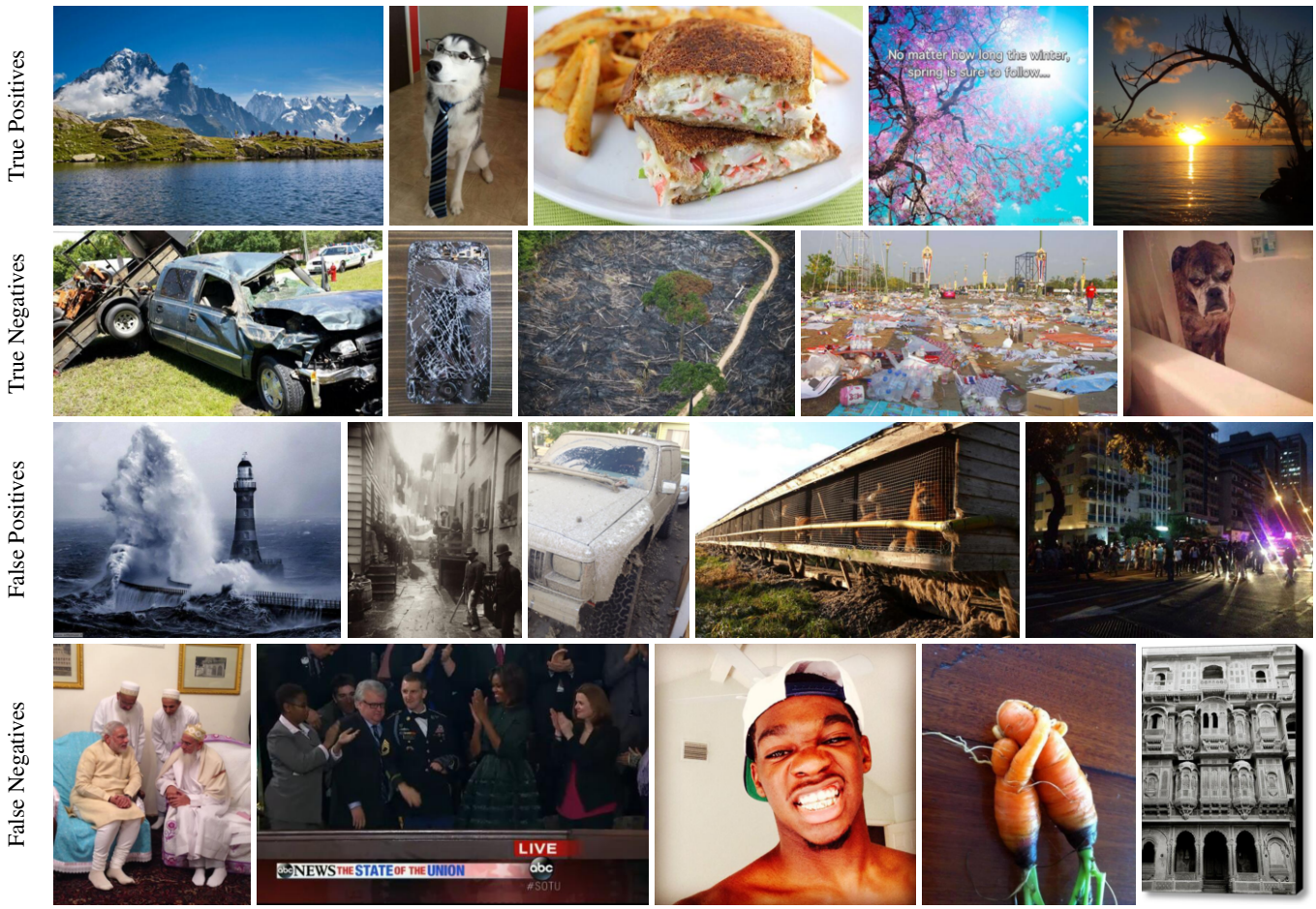

Figure 3. Cherry-picked examples of predictions of our best model (ViT-L/16) on the Twitter Dataset benchmark. The first two rows contain correctly classified images with positive and negative sentiment polarity, respectively. The third row contains negative-labeled samples misclassified as positives, and the last row contains positive-labeled ones misclassified as negatives. Note that several misclassified samples appear ambiguous even to a human labeler due to personal sensibility.

Table 4. Accuracy on standard benchmarks for visual-only image sentiment polarity prediction compared with state-of-the-art predictors.

	Twitter Dataset				
Model	5 agree	≥4 agree	≥3 agree	**Emotion ROI**	**FI**
Chen et al. [10]*	76.4	70.2	71.3	70.1	61.5
You et al. [43]*	82.5	76.5	76.4	73.6	75.3
Jou et al. [17]†	83.9±0.3				
Vadicamo et al. [40]	89.6	86.6	82.0		
Yang et al. [42]*	88.7	85.1	81.1	81.3	86.4
Wu et al. [41]	89.5	87.0	81.7	**83.0**	88.8
Ours (ViT-L/16, zero-shot)	87.8	84.8	81.9	64.1	76.0
Ours (ViT-L/16, fine-tuned)	**92.4±2.0**	**90.2±2.0**	**86.3±3.0**	**83.9±1.0**	**89.4±0.1**

*As reported by [41]. †As reported by [8].

extracted from Twitter data, applying the model as-is yielded a satisfactory performance only on same-domain data. Drawing data from a stream of multiple social media would improve zero-shot generalization and enable experimentation on larger scales. Moreover, confidence filtering is still manually tuned for the particular distribution of input data, while an adaptive online balancing of samples will be explored in future work.

Ethical Statement

Cultural and linguistic factors influence the predictions of our trained models that thus incorporate potential biases. This might be desirable when performing sentiment analysis on a specific population but could be detrimental when transferred to other communities without a thorough bias analysis.

Moreover, the use of sentiment analysis by large corporations to achieve commercial benefits poses an ethical issue, as it runs the risk of causing detrimental effects on individuals or groups of people. The proposed method is intended to be used in conjunction with ethical web scraping. The experiments reported in this work have been conducted exploiting the Twitter developer API complying with their Terms of Service.

Acknowledgements

This work was partially funded by: AI4Media - A European Excellence Centre for Media, Society and Democracy (EC, H2020 n. 951911); SUN - Social and hUman ceNtered XR (EC, Horizon Europe n. 101092612); SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by European Union - NextGenerationEU.

References

[1] Cosme Adrover, Todd Bodnar, Zhuojie Huang, Amalio Telenti, Marcel Salathé, et al., 'Identifying adverse effects of hiv drug treatment and associated sentiments using twitter', *JMIR public health and surveillance*, **1**(2), e4488, (2015).

[2] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, et al., 'Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.', in *Lrec*, volume 10, pp. 2200–2204, (2010).

[3] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke, 'Tweeteval: Unified benchmark and comparative evaluation for tweet classification', *arXiv preprint arXiv:2010.12421*, (2020).

[4] Adam Bermingham and Alan F Smeaton, 'Classifying sentiment in microblogs: is brevity an advantage?', in *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 1833–1836, (2010).

[5] Johan Bollen, Huina Mao, and Xiaojun Zeng, 'Twitter mood predicts the stock market', *Journal of computational science*, **2**(1), 1–8, (2011).

[6] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang, 'Large-scale visual sentiment ontology and detectors using adjective noun pairs', in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 223–232, (2013).

[7] Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al., 'TweetNLP: Cutting-Edge Natural Language Processing for Social Media', in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Abu Dhabi, U.A.E., (November 2022). Association for Computational Linguistics.

[8] Victor Campos, Brendan Jou, and Xavier Giro-i Nieto, 'From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction', *Image and Vision Computing*, **65**, 15–22, (2017).

[9] Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya, 'Context-aware interactive attention for multi-modal sentiment and emotion analysis', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5647–5657, (2019).

[10] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang, 'Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks', *arXiv preprint arXiv:1410.8586*, (2014).

[11] Nicholas A Diakopoulos and David A Shamma, 'Characterizing debate performance via aggregated twitter sentiment', in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1195–1198, (2010).

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., 'An image is worth 16x16 words: Transformers for image recognition at scale', *arXiv preprint arXiv:2010.11929*, (2020).

[13] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang, 'Image popularity prediction in social media using sentiment and context features', in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 907–910, (2015).

[14] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria, 'Misa: Modality-invariant and-specific representations for multimodal sentiment analysis', in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, (2020).

[15] Jyoti Islam and Yanqing Zhang, 'Visual sentiment analysis for social images using transfer learning approach', in *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom)*, pp. 124–130. IEEE, (2016).

[16] Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury, 'Twitter power: Tweets as electronic word of mouth', *Journal of the American society for information science and technology*, **60**(11), 2169–2188, (2009).

[17] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang, 'Visual affect around the world: A large-scale multilingual visual sentiment ontology', in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 159–168, (2015).

[18] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, 'Scaling laws for neural language models', *arXiv preprint arXiv:2001.08361*, (2020).

[19] Marie Katsurai and Shin'ichi Satoh, 'Image sentiment analysis using latent correlations among visual, textual, and sentiment views', in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2837–2841. IEEE, (2016).

[20] Zaid Khan and Yun Fu, 'Exploiting bert for multimodal target sentiment classification through input space translation', in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3034–3042, (2021).

[21] Aditya Khosla, Atish Das Sarma, and Raffay Hamid, 'What makes an image popular?', in *Proceedings of the 23rd international conference on World wide web*, pp. 867–876, (2014).

[22] Bing Li, Songhe Feng, Weihua Xiong, and Weiming Hu, 'Scaring or pleasing: exploit emotional impact of an image', in *Proceedings of the 20th ACM international conference on Multimedia*, pp. 1365–1366, (2012).

[23] Zuhe Li, Yangyu Fan, Weihua Liu, and Fengqin Wang, 'Image sentiment prediction based on textual descriptions with adjective noun pairs', *Multimedia Tools and Applications*, **77**, 1115–1132, (2018).

[24] Bing Liu and Lei Zhang, 'A survey of opinion mining and sentiment analysis', in *Mining text data*, 415–463, Springer, (2012).

[25] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han, 'On the variance of the adaptive learning rate and beyond', in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, (2020).

[26] Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados, 'Timelms: Diachronic language models from twitter', *arXiv preprint arXiv:2202.03829*, (2022).

[27] Jana Machajdik and Allan Hanbury, 'Affective image classification using features inspired by psychology and art theory', in *Proceedings*

of the 18th ACM international conference on Multimedia, pp. 83–92, (2010).

[28] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten, 'Exploring the limits of weakly supervised pretraining', in *Proceedings of the European conference on computer vision (ECCV)*, pp. 181–196, (2018).

[29] Yelena Mejova, Padmini Srinivasan, and Bob Boynton, 'Gop primary season on twitter: " popular" political sentiment in social media', in *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 517–526, (2013).

[30] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky, 'Distant supervision for relation extraction without labeled data', in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003–1011, (2009).

[31] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher, 'A mixed bag of emotions: Model, predict, and transfer emotion distributions', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 860–868, (2015).

[32] Plutchik Robert, 'Emotion: a psychoevolutionary synthesis', *New York7 Harper and Row*, (1980).

[33] Huaxia Rui, Yizao Liu, and Andrew Whinston, 'Whose and what chatter matters? the effect of tweets on movie sales', *Decision support systems*, **55**(4), 863–870, (2013).

[34] Mehmet Umut Salur and Ilhan Aydin, 'A novel hybrid deep learning model for sentiment classification', *IEEE Access*, **8**, 58080–58093, (2020).

[35] Stefan Siersdorfer, Enrico Minack, Fan Deng, and Jonathon Hare, 'Analyzing and predicting sentiment of images on the social web', in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 715–718, (2010).

[36] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta, 'Revisiting unreasonable effectiveness of data in deep learning era', in *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, (2017).

[37] Luam Catao Totti, Felipe Almeida Costa, Sandra Avila, Eduardo Valle, Wagner Meira Jr, and Virgilio Almeida, 'The impact of visual attributes on online image diffusion', in *Proceedings of the 2014 ACM conference on Web science*, pp. 42–51, (2014).

[38] Quoc-Tuan Truong and Hady W Lauw, 'Vistanet: Visual aspect attention network for multimodal sentiment analysis', in *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 305–312, (2019).

[39] Quoc-Tuan Truong and Hady W Lauw, 'Concept-oriented transformers for visual sentiment analysis', in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 1111–1119, (2023).

[40] Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi, 'Cross-media learning for image sentiment analysis in the wild', in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, (Oct 2017).

[41] Lifang Wu, Mingchao Qi, Meng Jian, and Heng Zhang, 'Visual sentiment analysis by combining global and local information', *Neural Processing Letters*, **51**(3), 2063–2075, (2020).

[42] Jufeng Yang, Dongyu She, Ming Sun, Ming-Ming Cheng, Paul L Rosin, and Liang Wang, 'Visual sentiment prediction based on automatic discovery of affective regions', *IEEE Transactions on Multimedia*, **20**(9), 2513–2525, (2018).

[43] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, 'Robust image sentiment analysis using progressively trained and domain transferred deep networks', in *Twenty-ninth AAAI conference on artificial intelligence*, (2015).

[44] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, 'Building a large scale dataset for image emotion recognition: The fine print and the benchmark', in *Proceedings of the AAAI conference on artificial intelligence*, volume 30, (2016).

[45] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu, 'Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis', in *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 10790–10797, (2021).