Facial Emotion Recognition for Sentiment Analysis of Social Media Data*

Diandre de Paula[0000−0002−6290−1440] and Luís A. Alexandre[0000−0002−5133−5025]

Departamento de Informática and NOVA LINCS, Universidade da Beira Interior, Covilhã, Portugal

Abstract. Despite the diversity of work done in the area of image sentiment analysis, it is still a challenging task. Several factors contribute to the difficulty, like socio-cultural issues, the difficulty in finding reliable and properly labeled data to be used, as well as problems faced during classification (e.g the presence of irony) that affect the accuracy of the developed models. In order to overcome these problems, a multitasking model was developed, which considers the entire image information, information from the salient areas in the images, and the facial expressions of faces contained in the images, together with textual information, so that each component complements the others during classification. The experiments showed that the use of the proposed model can improve the image sentiment classification, surpassing the results of several recent social media emotion recognition methods.

Keywords: Image Sentiment Analysis, Multimodal, Facial Expression Recognition, Salient Areas, Text Sentiment Analysis.

1 Introduction

We are increasingly witnessing the growth of the online community, where users seek ways to express themselves beyond the use of words, often using images to reach their goal. Thus, social media has posts with both text and images, that convey different (positive and negative) feelings. There are many factors to take into account when we analyze the sentiment transmitted by an image, for instance, the socio-cultural issues. Several other features can help us to identify the sentiment of an image, for example, the prevailing colors in the image, the type of objects in the image, and the metadata (e.g image's caption) that are associated with it. This work aims to develop a multimodal approach that classifies the image sentiment to identify posts that may represent negative and strongly negative situations, since we are interested in predicting when possible

* This work was supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT – Fundação para a Ciência e a Tecnologia, through national funds and by the project MOVES – Monitoring Virtual Crowds in Smart Cities (PTDC/EEI-AUT/28918/2017), also financed by FCT.

strongly negative events are going to take place, through the analysis of social media posts. This prediction will be obtained not just with the image information from the social media posts, but also with textual information. Several previous works have been done in this area [1–3, 9, 10], and we have identified a place where current models can be improved: the inclusion of a Face Emotion Recognition (FER) model can be used to clarify situations where the emotion conveyed by the text is not in agreement with the emotion in the image and also when the overall image has a positive sentiment if the face emotions are not taken into consideration. Besides this, we also experiment with the use of an image classifier for salient regions of the image, as to complement the information provided by a global image classier. Finally, we explore different ways to fuse the decisions from the proposed classifiers and present experiments on a large social media data set that show the strengths of our proposal.

2 Related Work

The work [10] approaches the image sentiment analysis (without considering text), using a Plutchik's wheel of emotions approach. It also addresses challenging issues like implementing supervised learning with weakly labeled training data, in other words, data that was labeled through a model (and not labeled by a human), and handles the image sentiment classification generalizability.

To predict the image sentiments, the authors of [9] proposed a model that combines global and local information. The work proposes a framework to leverage local regions and global information to estimate the sentiment conveyed by images, where the same pre-trained Convolutional Neural Network (CNN) model is used, but it is fine-tuned using different training sets: a first one addressing the entire images, and another addressing the sub-images; in the end, both predictions are fused to obtain the final sentiment prediction.

The work [2] aims to reduce the image classification's dependence on the text content. The proposed model was divided into three parts (in each one there is a specific task) and in the end, all parts are fused using a weighted sum, which is capable of predicting the polarity of a sentiment level (positive, neutral, and negative).

The authors in [1] propose a method based on a multi-task framework to combine multi-modal information whenever it's available. The proposed model contains one classifier for each task: i) text classification, ii) image classification, iii) prediction based on the fusion of both modalities. The authors evaluated the advantages of their multi-task approach on the generalization of each three tasks: text, image, and multi-modal classification. The authors concluded that their model is robust to a missing modality.

The work in [3] proposed a novel multi-modal approach, which uses both textual data and images from social media to perform the classification into three classes: positive, neutral, and negative. The approach consists of the classification of the textual and image components, followed by the fusion of both classifications into a final one using an Automated Machine Learning (AutoML)

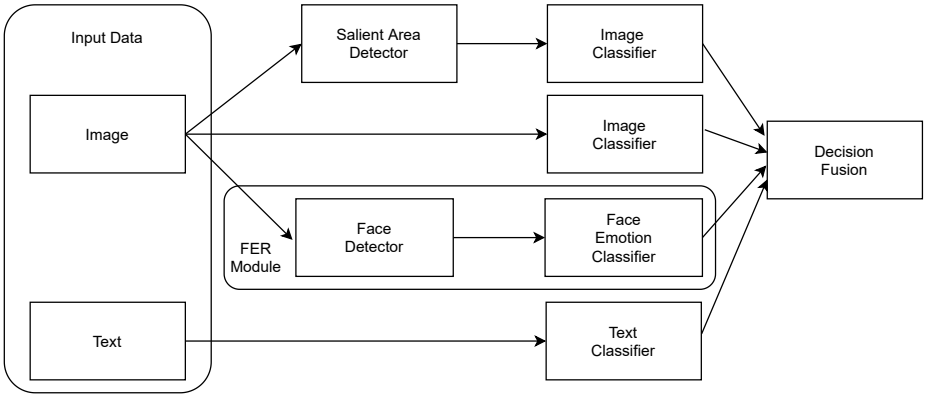

Fig. 1. Overview of the proposed method's architecture and its components.

approach, which performs a random search to determine the best model to perform the final classification.

We notice that, several works employed models with image and text classification. However, none of them employed an approach that could handle images, salient regions, textual data, and facial expressions. Regarding the final output, the majority of the work employed polarity as the final classification, using either two classes (positive and negative), or three classes (positive, neutral, and negative).

3 Proposed Method

Figure 1 shows an overview of the method that was developed. It's composed of an image classifier, an image salient area detector, a text classifier, and a facial expression module that contains both a face detector and a face emotion classifier. The outputs of the classifiers are fused to produce the final decision. The details are described in the next sub-sections.

3.1 Image Classifier

The image classifier model is responsible for analysing the sentiment of the original image. This is a mandatory model, that is, the information returned by this model will be always considered for the final fused decision. For the image classifier, the architecture proposed in [2] was used, since the proposed configurations obtained good results, better than the previous state-of-the-art [8]. A pre-trained Residual Network (ResNet) 152 was used, with the last layer being fully-connected, accompanied by a softmax layer, with 3 outputs, which represent the probability of each class (negative, neutral, and positive), in the range [0,1], where 1 represents that the image belongs to that respective class and 0

that it does not. The used model was trained with the B-T4SA data set (described in the Experiments section below). The model receives an image, and it predicts a class and its respective probability, which can be seen as the degree of certainty with which that class was predicted.

3.2 Salient Area Detector

The salient area detector is the component responsible for detecting the salient areas in the image. The objective of using a model that performs the detection of salient areas is to get a sense of which objects are contained in the images. Certain objects can strongly influence the sentiment of an image, like guns or other weapons, which negatively influence the sentiment, or flowers and beautiful landscapes, which influence the sentiment in a positive way. For a single image, several salient regions can be detected and our method considers only the one that has been detected with the highest confidence degree. The detector chosen was You Only Look Once (YOLO) v5 [7] which is a PyTorch implementation and includes mosaic data augmentation and auto-learning bounding box anchors. It provides four models: YOLO5S, YOLO5M, YOLO5L, and YOLO5X. We choose to use the largest model, YOLO5X, to aim for the best detection rates.

We used the VOC data set to train the detector. It contains 21,503 images with annotations. The data was split into 16,551 (77%) images for train and 4,952 (23%) for validation. The model was trained with a batch size of 64. The mean Average Precision (mAP)0.5 was 83.1%, the mAP@0.95 was 62.7% and the time it took to train the model was 27 minutes.

3.3 Text Classification

Since another important part of a social media post is the text, it also should be evaluated w.r.t. its sentiment.

We use the Regionbased Convolutional Neural Networks (R-CNN) text model proposed in [3], which uses an embedding layer, and a bi-directional Long Short Term Memory (LSTM) layer with input size equal to the dimension of the embedding, hidden size of 256 and a dropout of 0.8. The final embedding vector is the concatenation of its embedding and left and right contextual embeddings, which in this case is the hidden vector of the LSTM. This concatenated vector is then passed to a linear layer which maps the input vector back to a vector with a size equal to the hidden size of the LSTM, 256. This is passed through a 1D max-pooling layer, and finally, the output from this layer is sent to a linear layer that maps the input to a classification vector.

Tweets involve a lot of noise, such as emojis/emoticons, links, numbers, etc. Therefore, before going through the text evaluation method, the tweet must be cleaned, in order to remove this noise. First, the tweet will be processed by Beautiful Soup, which is a Python library for obtaining data from HyperText Markup Language (HTML) and eXtensible Markup Language (XML) files. It is used to decode HTML encoding that has not been converted to text, and ended

up in the text field, such as '&amp' or '&quot'. The second part of the preparation is dealing with @mention. Even though @mention carries some information (which user the tweet mentioned), this information does not add value to build a sentiment analysis model. The third part is dealing with Uniform Resource Locators (URL) links, that although they can carry some information for sentiment analysis purposes, they will be ignored. There is also the possibility of Unicode Transformation Format (UTF)-8 BOM character issues. The UTF-8 BOM is an array of bytes (EF BB BF) that allows the reader to recognize a file as being encoded in UTF-8. To avoid unfamiliar characters, we used a text decoder that replaces them by the symbol "?". Sometimes the text used with hashtags can give useful information about the tweet. So it was decided to leave the text intact and just remove the symbol (#) cleaning all the non-letter characters (including numbers). Then the text is transformed to lower case. During the letters-only process, unnecessary white space is created, so redundant white space is removed.

3.4 Facial Expression Recognition Module

Global features can give hints regarding the image sentiment, but some works faced difficulties with the models getting an erroneous classification due to global features [5, 6, 12]. Therefore, the objective of using a model that performs the classification of facial expressions would be to address these issues. Since in a single image we can have several faces, the information to be considered will be the one that has been obtained with the highest confidence degree. The Facial Expression Recognition Module is responsible for two tasks: i) detecting faces in the images; ii) classifying the detected faces' expressions. To make the detection, the Multi-task Cascaded Convolutional Networks (MTCNN) is used. This model has three convolutional networks (Proposal Network (P-Net), Refinement Network (R-Net), and Output Network (O-Net)). Upon receiving an image, the model will create an image pyramid, in order to detect faces of different sizes. Then, it is possible to split the MTCNN operation into three stages [11]:

- Stage 1: A fully convolutional network (P-Net) was used to obtain the candidate facial windows and their bounding box regression vectors. Candidates are calibrated based on the estimated bounding box regression vectors. Then, non-maximum suppression (NMS) is employed to merge highly overlapped candidates;
- Stage 2: All candidates are fed to another CNN (R-Net), which further rejects a large number of false candidates, performs calibration with bounding box regression, and conducts NMS;
- Stage 3: This stage is similar to the previous one, but it is aimed at identifying face regions with more supervision. In particular, the O-Net will output five facial landmarks' positions.

The FER model receives an image, and using the MTCNN, produces the bounding boxes together with the confidence degrees of each detected face in

the image. Then, if faces were detected in the image, each identified bounding box is cropped. The resolution of the cropped image will be checked in order to maintain a certain quality level of the image crops and prevent images of very low quality (and which would not add any utility to the model) from being kept. We used the rule of only keeping images with a resolution greater than or equal to 30x30 pixels.

After the face detector, the emotion recognition is performed. For this task, a CNN was created with a ResNet9, which increases (gradually) the number of channels of facial data and decreases the dimension, followed by a fully connected layer responsible for returning an array with the values describing the probability of belonging to each class. The learning rate scheduler, 1Cycle, was used so that the learning rate was not manually set. It starts with a very low learning rate, increases, and decreases it again.

3.5 Decision Fusion

To make the fusion of the information of each model, we propose two different methods: i) considering the average of all models and ii) using a voting system. To obtain the average of all models, each class obtained by the model and its respective accuracy will be multiplied, and this value will be divided by the number of models that were evoked, that is, to consider only the information of the models that were actually evoked:

$$av = \frac{1}{n} \sum_{i=1}^{n} X_i p_i \tag{1}$$

where n is the number of used models, X_i is the polarity and p_i is the accuracy values obtained by the global image, salient areas, text, and FER models, respectively in the validation set.

After obtaining the average result, the class is defined by:

$$class = \begin{cases} 0, & \text{if } av \leq 0.34 \\ 1, & \text{if } 0.34 < av \leq 0.67 \\ 2, & \text{if } av > 0.67 \end{cases} \tag{2}$$

the values 0.34 and 0.67 were chosen to divide the [0,1] interval into three equal intervals.

For the voting system, the votes for each class are counted, and to avoid any tie, the accuracy value of each model will be considered, when necessary. Therefore, a tuple is created, which stores the vote count of each class and the sum of the accuracy values of each model that voted in this same class:

$$(v_i, s_i), \quad i = 0, 1, 2, \tag{3}$$

where v_i is the vote count for class i and s_i is the sum of the accuracy values for that class. The selected class is given by:

$$class = \arg\max_i v_i \tag{4}$$

when there is no draw between the votes, and in the case of a draw, the class is given by:

$$class = \arg\max_i s_i, \tag{5}$$

where the index i runs through the drawn classes only. Then, to obtain the winner class, we will consider the index of the tuple with the highest value. If there is no tie, the tuple with the highest number of votes is chosen otherwise, the tuple (among those that are tied) with the highest sum is chosen.

4 Experiments

4.1 Training the Facial Expression Recognition Model

For training the FER model, we created a data set from three different sources.

First, the FER2013 data set was used. This data set consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is approximately centered and occupies the same amount of space in each image. The labels are divided into 7 types: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral. The training set consists of 28,709 examples and the validation and test sets consist of 3,589 examples each. The model achieved approximately 68.82% accuracy in the test set.

Another data set was prepared with social media images, namely from Twitter, in order to assess the model's behavior when exposed to social media images, which may or may not have larger resolutions. The accuracy obtained with the Twitter image test data set was approximately 18.22%.

The final data set used for training the FER model included all the images from the two previous data sets (FER2013 and Twitter) plus the Japanese Female Facial Expression (JAFFE) data set [4]. It contains 50,783 images, which were divided into: 40,627 samples for training (80%), 5,078 samples for testing (10%), and 5,080 samples for validation (10%). Figure 2 presents some of the images that compose the final data set used. The model was trained for 55 epochs, and the accuracy on the test set was 72.75%.

4.2 Data Set for the Full Model Evaluation

For the experiments, a variation of the B-T4SA validation set was used. In [8], the authors trained a model for visual sentiment classification starting from a large set of user-generated and unlabeled contents. They collected more than 3 million tweets containing both text and images. The authors used Twitter's Sample Application Programming Interface (API) to access a random 1% sample of the stream of all globally produced tweets, discarding tweets not containing any static image or other media, tweets not written in the English language,

Fig. 2. Sample images of the data set used for training the FER model. Sentiment from left to right (label, class, polarity): (surprise, 5, 1), (sad, 4, 0), (happy, 3, 2), (angry, 0, 0), and (neutral, 6, 1).

Table 1. Accuracy on the validation set (top 4 rows) and test set (last row) of several combinations of the available modules, with the training and evaluation times. We show the results for the fusion with the mean and voting approaches. Experiments A, B, C and D used the training and validation data, and the final experiment, E, was done using training and test data with the configuration that yielded the best results with the validation data.

Exp.	**Text clf.**	**Image clf.**	**Salient areas**	**FER clf.**	**Fusion Acc. [%]**		**Time [hours]**	
					Mean	**Vot.**	**Mean**	**Vot.**
A	Y	Y	N	N	60.22	-	0.30	-
B	Y	Y	Y	Y	59.31	73.19	9.15	9.15
C	Y	Y	Y	N	62.25	72.74	8.97	8.97
D	Y	Y	N	Y	62.63	**82.90**	0.51	0.51
E	Y	Y	N	Y	-	**80.86**	-	5.18

whose text was less than 5 words long, and retweets. At the end of the data collection process, the total number of tweets in the T4SA data set was about 3.4 million. Each tweet (text and associated images) was labeled according to the sentiment polarity of the text (negative=0, neutral=1, positive=2) predicted by their tandem LSTM-Support Vector Machines (SVM) architecture. The corrupted and near-duplicate images were removed, and they selected a balanced subset of images, named B-T4SA, that was used to train their visual classifiers.

The original B-T4SA validation set contains 51,000 samples. However, due to the hardware limitations, this set was randomly decreased to approximately 17% of its original size, resulting on a new validation set containing 9,064 samples.

Table 2. Comparison between the results obtained from the models that used the B-T4SA data set, all evaluated in the same test set.

Work	**Accuracy** [%]
VGG-T4SA FT-A [8]	51.30
VGG-T4SA FT-F [8]	50.60
Hybrid-T4SA FT-A [8]	49.10
Hybrid-T4SA FT-F [8]	49.90
Random Classifier [8]	33.30
Multimodal Approach [2]	52.34
Multimodal Approach [3]	**95.19**
Ours	80.86

4.3 Results

The experiments were run in a computer with an AMD Ryzen 7 2700 (Octacore, 16 Threads) 3.2GHz CPU, 16GB Random Access Memory (RAM), NVIDIA 1080ti, a 3TB HDD and a 256GB SSD.

We ran the first batch of experiments (corresponding to the first three rows of Table 1, experiments A, B, C and D) using only the training and validation data sets to study which was the best configuration in terms of model components to use with the final test data set, presented in the last row of Table 1, experiment E. The test data set contains 51,000 samples.

Test A was made in order to evaluate the model's accuracy without using the proposed methods. Since it only uses 2 models, the voting system was not used as the decision will be the same as if only the most accurate model was used.

From the validation set experiments, we found that the best results were achieved when not considering the salient region classifier, hence, in the final experiment E, with the test data, the configuration of our approach included only the text, global image and face emotion recognition modules. Regarding the two evaluated fusion approaches, the voting approach presented consistently over 10% better results than the fusion using the mean, as was the one evaluated in experiment E. Regarding the time that it took to train and evaluate the models, the table also shows that the salient area classifier was very costly and its removal significantly increase the speed of the system. The time increase from experiment D to E is due to the larger size of the test set when compared to the validation set.

We compared the results obtained in the test set with the results from other approaches in the literature that used the B-T4SA data set. The values are in Table 2 and show that our approach is able to improve on most of the previous results by a large margin, from around 50% to 80%, with the exception of the proposal in [3]. This points to the possibility of the module decision fusion used in that work, an AutoML approach, to be responsible for that large boost in accuracy, and to be a good alternative to the approaches explored in this paper.

5 Conclusions

Social media sentiment classification is a very demanding task. An approach that has been increasingly used is the analysis of both text and image (multi-modal) information to achieve improved results. In this paper, we also propose a multi-modal approach that uses text and image data from tweets to evaluate the post sentiment. We propose a system that explores the image data in several ways to try to overcome the ambiguity that can appear in the image sentiment evaluation. First, we use a global image classifier that is reused to process also salient image regions. From the experiments, we concluded that the salient regions' contribution to the final decision was not improving the overall classification results. We also proposed the use of a facial expression recognition (FER) module in the model. This approach has not been employed yet, or wasn't used with the three other models. The idea is to evaluate the emotion of the persons that might be present in the image and use this as a complement to the overall sentiment evaluation. The faces are detected and their emotions are obtained and we consider only the one with the highest confidence. To train the FER module, a data set was created, which contained images of faces in controlled environments and in the wild, in order to present a variety of possible situations, image quality and poses during the model's training. This module produced a good contribution to the final test set accuracy of the B-T4SA data set, of over 80%. The overall result largely improved previously obtained results with one exception only, that used AutoML to create a decision fusion from several models. In future work, we will study ways of improving our proposal that can compete with this approach.

References

1. Fortin., M., Chaib-Draa., B.: Multimodal sentiment analysis: A multitask learning approach. In: Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods - ICPRAM,. pp. 368–376. INSTICC, SciTePress (2019). https://doi.org/10.5220/0007313503680376
2. Gaspar, A., Alexandre, L.A.: A multimodal approach to image sentiment analysis. In: Intelligent Data Engineering and Automated Learning – IDEAL 2019. pp. 302–309. Springer International Publishing (2019)
3. Lopes, V., Gaspar, A., Alexandre, L.A., Cordeiro, J.: An automl-based approach to multimodal image sentiment analysis. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–9 (2021). https://doi.org/10.1109/IJCNN52387.2021.9533552
4. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, April **1998**, 200 – 205 (05 1998). https://doi.org/10.1109/AFGR.1998.670949
5. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: Proceedings of the 18th ACM International Conference on Multimedia. p. 83–92. MM '10, Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1873951.1873965, https://doi.org/10.1145/1873951.1873965

6. Ortis, A., Farinella, G.M., Battiato, S.: Survey on visual sentiment analysis. IET Image Processing **14**(8), 1440–1456 (May 2020). https://doi.org/10.1049/iet-ipr.2019.1270, http://dx.doi.org/10.1049/iet-ipr.2019.1270
7. Ultralytics: Yolov5, https://github.com/ultralytics/yolov5
8. Vadicamo, L., Carrara, F., Cimino, A., Cresci, S., Dell'Orletta, F., Falchi, F., Tesconi, M.: Cross-media learning for image sentiment analysis in the wild. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). pp. 308–317 (2017)
9. Wu, L., Qi, M., Jian, M., Zhang, H.: Visual sentiment analysis by combining global and local information. Neural Processing Letters **51**, 1–13 (06 2020). https://doi.org/10.1007/s11063-019-10027-7
10. You, Q., Luo, J., Jin, H., Yang, J.: Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. p. 381–388. AAAI'15, AAAI Press (2015)
11. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
12. Zhang, K., Zhu, Y., Zhang, W., Zhu, Y.: Cross-modal image sentiment analysis via deep correlation of textual semantic. Knowledge-Based Systems **216**, 106803 (2021). https://doi.org/https://doi.org/10.1016/j.knosys.2021.106803, https://www.sciencedirect.com/science/article/pii/S0950705121000666