A Multimodal Approach to Image Sentiment Analysis*

António Gaspar[0000−0002−6354−3374] and Luís A. Alexandre[0000−0002−5133−5025]

Universidade da Beira Interior
Instituto de Telecomunicações
Rua Marquês d'Ávila e Bolama, 6201-001,
Covilhã, Portugal
{antonio.pedro.gaspar, luis.alexandre}@ubi.pt

Abstract. Multimodal sentiment analysis is a process for the classification of the content of composite comments in social media at the sentiment level that takes into consideration not just the textual content but also the accompanying images. A composite comment is normally represented by the union of text and image. Multimodal sentiment analysis has a great dependency on text to obtain its classification, because image analysis can be very subjective according to the context where the image is inserted. In this paper we propose a method that reduces the text analysis dependency on this kind of classification giving more importance to the image content. Our method is divided into three main parts: a text analysis method that was adapted to the task, an image classifier tuned with the dataset that we use, and a method that analyses the class content of an image and checks the probability that it belongs to one of the possible classes. Finally a weighted sum takes the results of these methods into account to classify content according to its sentiment class. We improved the accuracy on the dataset used by more than 9%.

Keywords: Multimodal Sentiment Analysis · Image · Text · Deep Learning.

1 Introduction

Sentiment analysis is an important topic nowadays. With the advent of social media networks, the amount of data available is increasing exponentially which made sentiment analysis techniques and methods grow. These have many possible applications, among which are the anticipation of the behaviours and trends of the crowds. This kind of analysis is also used on opinion mining, which correlates the sentimental information with the influence of someone or something, that often has the purpose to convince or attract the individuals to do some

* This work was partially supported by Instituto de Telecomunicações under grant UID/EEA/50008/2019 and by project MOVES - Monitoring Virtual Crowds in Smart Cities (PTDC/EEI-AUT/28918/2017) financed by FCT - Fundação para a Ciência e a Tecnologia.

action, for example, to buy a new product or to vote in a determinate candidate for the elections. Although sentiment analysis is widely used for many tasks, its application has a high dependence on the textual content, which is present on the majority of comments that are published on social media networks.

Nonetheless and in spite of the textual content containing an objective message, which most of the times is clear to all of the participants when associated with an image, it can transform the image's natural meaning. A widely used expression is "A picture is worth a thousand words", meaning that an image can clearly transmit a message to the viewers that would otherwise require a large textual description to describe its contents. But unfortunately, the meaning of an image is not always clearly recognised because the viewers can have different cultures and life experiences, which means that they may have different ideas and perspectives about the interpretation of an image. This fact introduces subjectivity into the interpretation. Along this document we will present our proposal to resolve this challenging situation using current state-of-the-art methods on sentiment analysis, which are discussed in section 2. Our contribution is a method that can help to reduce the subjectivity of image sentiment analysis. The next section, presents the related work. Our method is presented in section 3. Section 4 contains the experiments and the last section the conclusions.

2 Related Work

2.1 Sentiments

In the psychology area, sentiments are different from emotions as is described by the authors of the papers [9, 6]. Sentiments are the result of subjective experiences that were lived from an emotion. Emotions, in general, are the triggers for actions that can be positive or negative. An emotion can occur as a response to an internal or external signal of the environment context. For example, pain, which is considered an internal signal, can trigger in most of cases the sadness emotion that produces bad feelings. Playing a game, can trigger joy and pleasure which are positives feelings. Emotions are the base of sentiments. These can construct the history of the all feelings that are processed and memorised. This fact is important in sentiment analysis because through it is possible to reduce the subjectivity according to the culture where the analysed data belongs to. However, often the data cannot be organised by the culture. It is the case of the data collected from social media networks. For this reason, artificial intelligence may help to find the best features for classification. Next, we present some of these techniques related with the present theme.

2.2 Sentiments and Artificial Intelligence

Nowadays text, images, videos and all multimedia content can be processed and analysed. This process is supported essentially by models that run on computers. These models can obtain important information from raw data. To analyse the

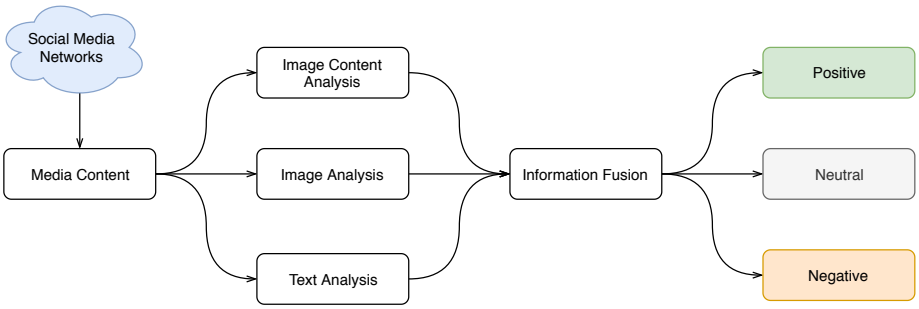

Fig. 1: Overview of the proposed method.

data, most models represent information using sets of features which in turn represent the classes of the target objects. This process can be done through many different approaches, but currently, deep neural networks such as Convolutional Neural Networks (CNNs) have been producing very good results when applied to image data.

There have been many proposals of methods for Image Sentiment Analysis. The authors of [10] studied the sentiment analysis process. They propose a method that is capable of classifying images at sentiment polarity level. The dataset they use is composed of 3 million tweets, which include text and images, and was constructed by them collecting the information on Twitter. For the classification, they propose a method that leverages the text classification and correlates it with the image. They conclude that text associated with image is often noisy and is weakly correlated with the image content, but it is possible to classify its sentiment using a model that is trained with the images classified with text labels. In another work described in [4], the authors explore four different architectures of convolutional neural networks to do sentiment analysis in visual media. This work was based on a labelled set that has the main categories of the description of the scene. With their results, the authors compose their own dataset and train a model that improves the results.

3 Proposed Method

Our method is based on a multimodal approach. It is composed of three parts that are fused in the end. Each part is designed for a specific task. The first part corresponds to the text analysis and the second part corresponds to the isolated image analysis. Both parts classify the sentiment analysis polarity, which is negative, neutral and positive. Next, in the third part, we studied the most common object class occurrence on the training and validation dataset, for each polarity class. Finally, we fuse all the parts using a weighted sum that is capable of predicting the polarity at the sentiment level. Figure 1 contains an overview of the proposed method.

3.1 Text Classification

Text sentiment analysis is a procedure derived from natural language processing (NLP). There are many proposed methods in this field, that are able to handle the job. These can be applied to problems like opinion mining and crowd influence through social networks. In this work we test two current state-of-the-art methods, the Vader [7] and the TextBlob [3] methods.

The Vader method (Valence Aware Dictionary and Sentiment Reasoner) [7] is a method composed by a list of lexical features, that are labelled according to their semantic orientation. Vader can produce four classifications, which are, negative, neutral, positive and compound score. Vader does not require to be trained because it is constructed through a standard sentiment lexicon.

The TextBlob [3] is a Python library that implements methods for processing textual data. It provides an API for processing NLP tasks such as part-of-speech tagging, and sentiment analysis. The sentiment analysis polarity, has a float range between -1 and 1, where values above 0.1 means positive, values below -0.1 means negative and values between -0.1 and 0.1 means neutral.

For the text analysis process, we tested both methods on the dataset described in section 4.1. To compare both methods, we use the respective confusion matrices that take into consideration the results that the *B-T4SA* dataset already provides on the validation set. With the analysis of these results, the TextBlob method is the chosen since it is the one that reveals a higher accuracy (64.271% vs 41.078%) in the *B-T4SA* dataset.

3.2 Image Classification

The developed method for image analysis is based on a deep learning approach. This is implemented with Pytorch [8], which is a deep learning framework that supports several features and automatic differentiation. For this work, we explore three versions of the Resnet, which are, the ResNet18, the ResNet50, and the ResNet152. We use the ResNet topology because it is a state-of-the-art method that reduces significantly the vanishing gradient problem. To use these models we need to set them up and prepare the data. To do that we follow the next steps.

1. **Data Preprocessing**: One of the biggest challenges on deep learning approaches is the data quality. Any deep learning approach is hungry for data, because it is through it that the network extracts and learns the features used for classification. The dataset used has many images with different scales and sizes. This fact can slow the training process. The pre-processing method used scales and re-sizes each image to 224x224.

2. **Model Choice and its Adaptation**. Pytorch comes with several built-in models. In this work, we selected three of these models, the ResNet18, ResNet50, and ResNet152. All the models are set up with the same hyper-parameters. These are, the learning rate that starts with 0.001, the momentum with 0.9 and the gamma parameter with 0.1. We use an optimiser that

will hold the current state and will update the parameters based on the computed gradients. This is the SGD (Standard Gradient Descendent). We use a schedule that provides several methods to adjust the learning rate based on the number of epochs. This will adjust the learning rate in every seven epochs. We define 30 epochs to train.

3. **Model Training and Evaluation** We train the models with a GeForce GTX 1080 TI, using the training set to train and the validation set to validate the training phase. ResNet18 uses 512 features from each image and achieves 50.3%, ResNet50 and ResNet152 use 2048 features and achieve better results. ResNet152 achieves the best result, 52.2%, and exceeds the result presented in the dataset paper [10] for only the image analysis, which is 51.3%. For this reason this is the model selected for the next phase.

3.3 Image Content Analysis

Image content analysis is a complex subject because an image might contain many objects. In this work we try to identify automatically the class of the object that an image can represent. To do this we use a pre-trained model with the ImageNet [5] to classify the data into its class through the ImageNet classifications (1000 possible object classes). All images on the ImageNet are quality-controlled and human-annotated. We use an InceptionResNetV2 trained model, which according to the author [1], has 80.17% of accuracy. This model comes from a python package that is called pretrainedmodels [2]. In this work, our intention with the image content analysis is to build a probability distribution that makes it possible to classify an image according to its sentiment polarity, such as negative, neutral and positive. So, with the InceptionResNetV2, we built a model that was feed with the union of the training and validation sets to increase the number of the images. The InceptionResNetV2 classified the contents of each dataset image into one of the ImageNet classes. Each of these images contains a sentiment classification in the training and validation sets that we used to build a table with the probability distribution of the image sentiment for each ImageNet class. In the table 1 we present an example of the full table (1000 rows). With this analysis, we can increase the information that we give to our proposal method. Next, we present the fusion of these three approaches to build our proposed method.

Table 1: An example of the results of our method with ImageNet.

Class ID	Class Name	Negative	Neutral	Positive
445	bikini, two-piece	27.27%	33.02%	39.70%
700	paper towel	40.54%	27.03%	32.43%
966	red wine	27.59%	20.69%	51.72%

3.4 Information Fusion

We built a method where we join the three methods explained before. This is used to classify multimedia content at sentiment polarity level, without a high text dependence. To do that, we make a weighted sum where we attribute the normalised validation accuracy value of each method as a weight, which is used to balance the importance that we give to each method.

The accuracy in the validation set of the three methods was 64.27, 52.15 and 40.09, respectively. The sum of these values is 156.51 and it is used to normalise them and create the weights used for producing the final decision:

$$w_1 = 0.41, w_2 = 0.33, w_3 = 0.26.$$

The results of each method are now used in a voting system. Each of the three partial methods vote on its own decision class with a value equal to its normalised weight, that is, the text-based decision votes with 0.41, the image-based with 0.33 and the image content-based with 0.26. The image content-based decision is simply the sentiment that is more probable for the given image class. The final decision is the class that has the highest vote value.

4 Experiments

In this section, we present the dataset used to evaluate our method, as well as the results that we achieved in this work. The results that we present, were obtained on the test set.

4.1 The Dataset

The authors of [10] built a dataset with three million tweets. These tweets contain text and images. Nonetheless this huge amount of data, it has some problems, such as, duplicated entries and malformed images. Another problem is the occurrence number around the three possible classes, negative, neutral and positive. In this case, positive and neutral content occur more times than the negative content. These situations led the authors to build a subset that is composed by tweets that have images and text in their corpus, non duplicated and non malformed images, as well the same number of occurrences on the different classes.

The subset is called *B-T4SA*, and is divided into three partitions: the train part, the validation part and the test part. Each one of these subsets has three classes, negative, neutral and positive. Each class has the same number of images as the others. Figure 2 shows some example tweets of the three classes.

4.2 Results

The result which we obtained is 60.42% accuracy on the test set. Table 2 shows the confusion matrix of our method and table 3 presents the results of the baseline and of our method. We can conclude that, with this results, it is possible

(a) **Negative**: "Not only are we surrounded by clowns, fools and useful idiots,we now have 000s more rapists, parasites and predators."

(b) **Neutral**: "Moon Mission: Hakuto robot rover undergoes testing in Japan ahead of lunar travel."

(c) **Positive**: "Join us tomorrow for another incredible Sunday! We have three services catering for all."

Fig. 2: Examples of the three classes present in the dataset, negative, neutral and positive.

Table 2: Confusion matrix of the proposed method in the test set. The accuracy in the test set is 60.42%

		Predicted Class		
		Negative	Neutral	Positive
True Class	Negative	8841	4960	3199
	Neutral	2980	9998	4022
	Positive	2338	2689	11973

to classify multimedia content using our method. Nonetheless there is space to improve it, for instance, by using all of the probabilistic information gathered for the image content-based decision instead of using only the most probable class.

5 Conclusions

In this work, we explore the sentiment analysis of tweets that contain both text and image, focusing on images and their content. We achieve a result on the isolated method image that exceeds the baseline method for the same theme in the paper [10]. We built a probability distribution table, that is based on 1000 classes of the ImageNet, that summarise a probability of a given image being negative, neutral or positive according to its content. Finally, we built a method that can classify multimedia content with text and image and generate a sentiment classification based on the image content. This method improves the results presented in [10] by 9%. For future work we intend to further improve the method and make more tests with other datasets.

Table 3: Results comparison between the method by the authors of the baseline paper [10] and the proposed method.

Method [10]	**Proposed Method**
51.30%	60.42%

References

1. Pretrained Models GitHub pretrained models for pytorch github. https://github.com/cadene/pretrained-models.pytorch, accessed: 2019-06-17
2. Pretrained Models pretrained models for pytorch. https://pypi.org/project/pretrainedmodels/, accessed: 2019-06-17
3. TextBlob textblob. https://textblob.readthedocs.io/en/dev/, accessed: 2019-06-17
4. Bonasoli, W., Dorini, L., Minetto, R., Silva, T.: Sentiment analysis in outdoor images using deep learning pp. 181–188 (10 2018). https://doi.org/10.1145/3243082.3243093
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
6. Hovy, E.H.: Language Production, Cognition, and the Lexicon **48**, 13–25 (2015). https://doi.org/10.1007/978-3-319-08043-7, http://link.springer.com/10.1007/978-3-319-08043-7
7. Hutto, C., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text (01 2015)
8. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
9. Pawar, A.B., Jawale, M.A., Kyatanavar, D.N.: Sentiment Analysis and Ontology Engineering **639** (2016). https://doi.org/10.1007/978-3-319-30319-2, http://link.springer.com/10.1007/978-3-319-30319-2
10. 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 (Oct 2017). https://doi.org/10.1109/ICCVW.2017.45