An Overview on Image Sentiment Analysis: Methods, Datasets and Current Challenges

Alessandro Ortis[a], Giovanni Maria Farinella[b] and Sebastiano Battiato[c]
University of Catania, Department of Mathematics and Computer Science, Viale A. Doria, 6, Catania, 95125, Italy

Keywords: Image Sentiment Analysis, Social Image Sentiment, Image Polarity Classification.

Abstract: This paper introduces the research field of Image Sentiment Analysis, analyses the related problems, provides an in-depth overview of current research progress, discusses the major issues and outlines the new opportunities and challenges in this area. An overview of the most significant works is presented. A discussion about the related specific issues is provided: emotions representation models, existing datasets and most used features. A generalizable analysis of the problem is also presented, by identifying and analyzing the components that affect the sentiment toward an image. Furthermore, the paper introduces some additional challenges and techniques that could be investigated, proposing suggestions for new methods, features and datasets.

1 INTRODUCTION

With the rapid diffusion of social media platforms (i.e., social networks, forums, and blogs), political parties, companies, and other organizations are increasingly taking into account public opinions for their business strategies (Liu and Zhang, 2012). The analysis of such social information enables applications such as brand monitoring, market prediction or political voting forecasting. In particular, companies are interested in monitoring people opinions toward their products or services (e.g., brand monitoring), by the other end, customers consult other users' feedbacks to assess the quality of a service or a product. The basic task in Sentiment Analysis is the polarity classification of an input text in terms of positive, negative or neutral polarity. The techniques in this field are used to infer public opinions. As an example, the paper in (Tumasjan et al., 2010) shown that the populairty of political parties can be predicted from the Twitter's users activity, as Twitter posts reflect the political landscape. So far, most of the efforts in Sentiment Analysis have been devoted to the analysis of text (Pang and Lee, 2008), whereas limited efforts have been employed to infer sentiments from images and videos, which are becoming a pervasive media on the web.

2 STATE OF THE ART

This section presents an overview of the most significant works in the field of Image Sentiment Analysis published since 2010, when the first significant work on the field that aims to classify images as "positive" or "negative" has been presented (Siersdorfer et al., 2010).

Most of the works in Image Sentiment Analysis are based on previous studies on emotional-aware image retrieval (Colombo et al., 1999; Schmidt and Stock, 2009), which try to find connections between emotions and image visual features, aimed to perform image retrieval and classification. In (Siersdorfer et al., 2010) the authors exploited meta-data associated to images to assign sentiment scores to each picture, and studied the correlations between the sentiment of images, in terms of positive or negative polarity, and visual features (e.g., local/global colour histogram). The authors of (Machajdik and Hanbury, 2010) studied the contribute of visual features for the task of affective classification of images. The image features have been selected by exploiting experimental observation and insights of emotional responses with respect to colours. The image classification considered 8 emotional classes defined in (Mikels et al., 2005), namely: Anger, Awe, Amusement, Excitement, Contentment, Disgust, Fear and Sad. A large scale Visual Sentiment Ontology (VSO) of semantic concepts based on psychological theories and web mining has been presented in (Borth et al., 2013). A

[a] https://orcid.org/0000-0003-3461-4679
[b] https://orcid.org/0000-0002-6034-0432
[c] https://orcid.org/0000-0001-6127-2470

Ortis, A., Farinella, G. and Battiato, S.
An Overview on Image Sentiment Analysis: Methods, Datasets and Current Challenges.
DOI: 10.5220/0007909602900300
In *Proceedings of the 16th International Joint Conference on e-Business and Telecommunications (ICETE 2019)*, pages 290-300
ISBN: 978-989-758-378-0

concept is expressed as an adjective-noun combination called Adjective Noun Pair (ANP). The authors also trained a set of 1.200 visual concept detectors (i.e., one for each ANP) which outputs can be exploited as a mid-level representation for an input image. The authors of (Chen et al., 2014) fine-tuned a Convolutional Neural Network (CNN) previously trained for the task of object classification to classify images in one of the 2.096 ANP categories (obtained by extending SentiBank (Borth et al., 2013)), the resulting CNN model is known as DeepSentiBank. In (Xu et al., 2014) the authors exploited a pre-trained CNN to extract two high-level features (i.e., the *fc7* and *fc8* activations), and then trained two Logistic Regressors on the task of sentiment classification. The performed sentiment classification considered 5 sentiment polarity classes: strong negative, weak negative, neutral, weak positive, and strong positive. A progressive CNN (PCNN) training approach has been implemented in (You et al., 2015), in order to perform binary sentiment image classification. First, a CNN architecture has been trained with a dataset of $\sim 0.5M$ Flickr images (Borth et al., 2013). During the training, a subset of training images which achieve high prediction scores are selected. Then, this subset is used to further fine-tune the obtained CNN. Differently than previous approaches, the authors of (Peng et al., 2015) aimed to predict a distribution representation of the emotions rather than a single emotion category. An interesting result of (Peng et al., 2015) is that the authors are able to change the colours and the textures of a given image according to the ones of a target image. The resulting image is able to evoke emotions closer to the target image than the original version. Such experiments suggest that colour tone and textures can influence the evoked emotion, and can be exploited to change the sentiment evoked by an image, beside its semantic content. In (Campos et al., 2015) a CNN pre-trained for the task of Object Classification is fine-tuned on the image sentiment prediction task. Then, with the aim to understand the contribution of each CNN layer for the task, the authors performed an exhaustive layer-per-layer analysis of the fine-tuned model.

The authors of (Sun et al., 2016) proposed an algorithm that combines features from either the whole image (i.e., global features) and dominant objects located in salient regions. The work in (Katsurai and Satoh, 2016) defines an embedding space in which the correlation between the projections of textual and visual features maximized. The projections in the defined embedding space are used to train a Support Vector Machine classifier on the task of binary image sentiment classification.

In (Yang et al., 2017) the authors proposed two Conditional Probability Neural Networks (CPNN), called Binary CPNN (BCPNN) and Augmented CPNN (ACPNN), for the task of emotional image classification. A CPNN is an Artificial Neural Network with one hidden layer which takes both features and labels as input and outputs the label distribution. Indeed, the aim of a CPNN is to predict the probability distribution over a set of considered labels. The authors of (Jufeng Yang, 2017) changed the dimension of the last layer of a CNN pre-trained for Object Classification in order to extract a probability distribution with respect to the considered emotional categories, and replaced the original loss layer with a function that integrates the classification loss and sentiment distribution loss through a weighted combination. Then the modified CNN has been fine-tuned to predict sentiment distributions. In (Campos et al., 2017), the activations in each layer of the CNN presented in (Campos et al., 2015) are used to train a layer-specific classifier. The work further presents a study on the effect of weight initialization in the context of transfer learning. The insights of this experimental study have been then exploited to further improve the classifier. The authors of (Vadicamo et al., 2017) built a dataset of $\sim 3M$ tweets containing text and images to finetune a CNN previously trained for objects and places classification (VGG19 (Simonyan and Zisserman, 2014) and HybridNet (Zhou et al., 2014)).

The work in (Ortis et al., 2018) addressed the problem of the noise in the text associated to social images by users, due to their subjectivity. In particular, the authors built an image sentiment classifier that exploits a representation based on both visual and textual features extracted from an image, and evaluated the performances obtained by using the text provided by users (subjective) and the text extracted from the visual content by using four deep models trained on different tasks. The experiments revealed that employing a source of text automatically extracted from the images in lieu of the text provided by users improves the classifier performances.

The works above have address the problem of Image Sentiment Analysis by considering different emotion models (i.e., classification schemes), datasets and evaluation methods. So far, researchers formulated this task as a classification problem among a number of polarity levels or emotional categories, but the number and the type of the emotional outputs adopted for the classification are arbitrary (see Table 1). The differences in the adopted system design components (i.e., emotional model, dataset and features) make result comparison difficult. Moreover, there is not a strong agreement in the research community about the use of an universal benchmark dataset. Many of the works presented in this section have at least one of the above described issues.

3 SYSTEM DESIGN COMPONENTS

This section provides a complete overview of the system design choices, with the aim to provide a comprehensive debate about each specific issue, with proper references to the state of the art.

3.1 Emotion Models

The basic task of an Image Sentiment Analysis system is to predict the sentiment polarity of an input image in terms of two (positive, negative) or three polarity levels (positive, neutral, negative). Nevertheless, there are approaches that adopt more then three levels, such as the 5-level sentiment categorization used by Xu et al. (Xu et al., 2014). Besides, there are systems that perform the sentiment classification by using a set of emotional categories, according to an established emotion model based on previous psychological studies, such as the 35 impression words defined in Hayashi et al. (Hayashi and Hagiwara, 1998). Each emotional category corresponds to a positive or negative polarity (Machajdik and Hanbury, 2010). Therefore, these approaches can be employed for the task of polarity estimation. In general, there are two main approaches for emotion modelling: The *Dimensional Model* represents emotions as points in a two or three dimensional space. Indeed, as discussed in several studies (Bradley, 1994; Lang, 1993; Osgood, 1952; Russell and Mehrabian, 1977), emotions have three basic underlying dimensions, namely the valence, the arousal and the control (or dominance) dimensions. However, the control dimension has a small effect. Therefore, a 2D emotion space is often considered. This space is obtained by considering only the arousal and the valence axis.

According to the *Categorical Model*, there are a number of basic emotions. This set of descriptive words can be assigned to regions of the VAC (Valence-Arousal-Control) space. Thus it can be considered a quantized version of the dimensional approach. There are several works that aim to define the basic emotions, as the choice of the emotional categories in not an easy task. The most adopted model is the Plutchnik's wheel of emotions (Plutchik, 1980) that defines 8 basic emotions with 3 valences each. According to Ekman's theory (Ekman et al., 1987) there are just five basic emotions ("anger", "fear", "disgust", "surprise" and "sadness"). Another emotions categorization is the one defined in a psychological study in (Mikels et al., 2005), where the authors perform an intensive study on the *International Affective Picture System* (IAPS) in order to extract a categorical structure of such dataset (Lang et al., 1999). As result, a subset of IAPS have been categorized in eight distinct emotions: "amusement", "awe", "anger", "contentment", "disgust", "excitement", "fear" and "sad". A deeper list of emotion is described in Shaver et al. (Shaver et al., 1987), where emotion concepts are organized in a hierarchical structure.

3.2 Image Sentiment Analysis Datasets

In the context of Sentiment Analysis the collection of huge opinion data can be obtained by exploiting the most common social platforms (Instagram, Flickr, Twitter, Facebook, etc.), as well as websites for collecting business and products reviews (Amazon, Tripadvisor, Ebay, etc.). Indeed, nowadays people are used to express their opinions and share their daily experiences through the Internet Social Platforms. Such platforms can be exploited to create specific dataset at large scale with very low effort. One of the most important aspect of this way to build datasets is that the data comes from real users. This paradigm is known as crowdsourcing, and the main advantage of this approach is that the data reflects preferences, behaviours and interactions of real users that publish, share and comment contents through the main social platforms (Battiato et al., 2016).

One of the first public dataset related to the task of Image Sentiment Analysis is the International Affective Picture System (IAPS) (Lang et al., 1999). Such dataset has been developed with the aim to produce a set of evocative colour images that includes contents from a wide range of semantic categories. This work provides a set of standardized stimuli for the study of human emotional process. In (Yanulevskaya et al., 2008) the authors considered a subset of IAPS extended with subject annotations to obtain a training set categorized in distinct emotions according to the emotional model described in (Mikels et al., 2005). However, the number of images of this dataset is very low. In (Machajdik and Hanbury, 2010), the authors presented the Affective Image Classification Dataset. It consists of two image sets: one containing 228 abstract painting and the other containing 807 artistic photos. These images have been labelled by using the 8 emotions defined in (Mikels et al., 2005). The authors of the dataset presented

Table 1: Summary of the most relevant publications on Image Sentiment Analysis.

Year	Paper	Input	Output
2010	(Siersdorfer et al., 2010)	Hand crafted visual features	Sentiment Polarity
2010	(Machajdik and Hanbury, 2010)	Hand crafted visual features	Emotional Classification (Mikels et al., 2005)
2013	(Borth et al., 2013)	ANP output responses	Sentiment Polarity and Emotional Classification (Mikels et al., 2005)
2014	Chen et al. (Chen et al., 2014)	Raw image	ANP annotation (Borth et al., 2013)
2014	(Xu et al., 2014)	CNN activations	5-scale Sentiment Score
2015	(You et al., 2015)	Raw image	Sentiment Polarity
2015	(Peng et al., 2015)	Hand crafted visual features	Emotions Distribution
2015	(Campos et al., 2015)	Raw image	Sentiment Polarity
2016	(Sun et al., 2016)	Image salient regions	Sentiment Polarity
2016	(Katsurai and Satoh, 2016)	Hand crafted visual features & textual metadata	Sentiment Polarity
2017	(Yang et al., 2017)	Raw image	Emotions Distribution
2017	(Jufeng Yang, 2017)	Raw image	Emotions Distribution
2017	(Campos et al., 2017)	Raw image	Sentiment Polarity
2017	(Vadicamo et al., 2017)	Raw image	Sentiment Polarity
2018	(Ortis et al., 2018)	Hand crafted visual features & text extracted from the image	Sentiment Polarity

in (Siersdorfer et al., 2010) considered the top 1.000 positive and negative words in SentiWordNet (Esuli and Sebastiani, 2006) as keywords to search and crawl over 586.000 images from Flickr. The Geneva Affective Picture Database (GAPED) (Dan-Glauser and Scherer, 2011) dataset includes 730 pictures labelled considering negative (e.g., images depicting human rights violation scenes), positive (e.g., human and puppies) as well as neutral pictures which show static objects. All dataset images have been rated considering the valence, arousal, and the coherence of the scene. The dataset is available for research purposes [1]. In (Borth et al., 2013) the authors proposed a very large dataset composed by $\sim$ 0.5 million pictures gathered from social media (i.e., from Flickr) and labelled with ANP (Adjective Noun Pair) concepts. Furthermore, they proposed a Twitter benchmark dataset which includes 603 tweets with photos. Both datasets and related annotations are publicly available. The dataset built in (Borth et al., 2013) has been used by most of the state-of-the-art works as evaluation benchmark for Image Sentiment Analysis, especially when the designed approaches involve the use of Machine Learning methods such as in (Yuan et al., 2013) and in (You et al., 2015) for instance, due to the large scale of this dataset. The work in (Peng et al., 2015), presented the Emotion6 dataset, built considering the Elkman's 6 basic emotion categories (Ekman et al., 1987). The number of images is balanced over the considered categories and the emotions associated with each image is expressed as a probability distribution instead of as a single dominant emotion. In (You et al., 2015) the authors proposed a dataset with 1,269 Twitter images labelled into positive or negative by five different annotators. In (Katsurai and Satoh, 2016) two large sets of social pictures from Instagram and Flickr (CrossSentiment) have been crawled. The list of labelled Instagram and Flickr image URLs is available on the Web [2]. In (Vadicamo et al., 2017) the authors crawled $\sim 3M$ tweets from July to December 2016. The collected tweets have been filtered considering only the ones written in English and including at least an image. The sentiment of the text extracted from the tweets has been classified using a polarity classifier based on a paired LSTM-SVM architecture. The data with the most confident prediction have been used to determine the sentiment labels of the images in terms of positive, negative and neutral. The resulting Twitter for Sentiment Analysis dataset (T4SA) consists of $\sim 1M$ tweets and related $\sim 1.5M$ images.

Table 2 reports the main details of the described datasets, including the source and number of the included images, the available annotation, if the images have been crawled from social media platforms, if the dataset has been labelled for the polarity classification task, and if additional meta-data (e.g., post and/or author information) is included in the dataset. Furthermore, for each dataset, the download URL is provided.

3.3 Features

One of the most important step driving the design of a Image Sentiment Analysis system, and in general for the design of a data analysis approach is the selection of the data features that are supposed to encode the information that the system is aimed to infer. Several studies have been conducted to assess the corre-

[1] http://www.affective-sciences.org/home/research/materials-and-online-research/research-material/

[2] http://mm.doshisha.ac.jp/senti/CrossSentiment.html

Table 2: Main benchmark datasets for Image Sentiment Analysis. Some datasets contains several additional information and annotations.

Dataset and URL	Size	Labelling	Social Media	Polarity	Additional Metadata
IAPS (Lang et al., 1999) https://csea.phhp.ufl.edu/media.html	716 photos	Pleasure, arousal and dominance	✗	✗	✗
(Mikels et al., 2005) https://link.springer.com /article/10.3758/BF03192732	369 photos	Awe, amusement, contentment, excitement, disgust, anger, fear, sad	✗	✗	✗
Affective Image Classification Dataset (Machajdik and Hanbury, 2010) http://www.imageemotion.org/	228 paintings 807 photos	Awe, amusement, contentment, excitement, disgust, anger, fear, sad	✗	✗	✗
Flickr-sentiment (Siersdorfer et al., 2010) http://www.l3s.de/~minack/flickr-sentiment/	586.000 Flickr photos	Positive, negative.	✓	✓	✓
GAPED (Dan-Glauser and Scherer, 2011) https://www.unige.ch/cisa/index.php /download_file/view/288/296/	730 pictures	Positive, negative, neutral.	✗	✓	✗
VSO (Borth et al., 2013) https://visual-sentiment-ontology.appspot.com/	0,5 M Flickr Photos 603 Twitter Images	- Adjective-Noun Pairs - Positive or negative	✓	✓	✓
Emotion6 (Peng et al., 2015) http://chenlab.ece.cornell.edu/downloads.html	1.980 Flickr photos	- Valence-Arousal score - 7 emotions distribution	✓	✗	✗
(You et al., 2015) https://www.cs.rochester.edu/u/qyou/ DeepSent/deepsentiment.html	1.269 Twitter images	Positive, negative.	✓	✓	✗
CrossSentiment (Katsurai and Satoh, 2016) http://mm.doshisha.ac.jp/senti/ CrossSentiment.html	90.139 Flickr photos 65.439 Instagram images	Positive, negative, neutral.	✓	✓	✗
T4SA (Vadicamo et al., 2017) http://www.t4sa.it/	1,5 M Twitter images	Positive, negative, neutral.	✓	✓	✗

lation between low, mid and high level visual features with the emotional effect of an image. One example is given by the scene-based 102-dimensional feature defined in (Yuan et al., 2013). Furthermore, many of the aforementioned works on Image Sentiment Analysis exploit the 1200-dimensional mid-level representation given by the 1200 Adjective-Noun Pairs (ANP) classifiers defined by Borth et al. (Borth et al., 2013).

In (Machajdik and Hanbury, 2010) an intensive study on image emotion classification by properly combining the use of several low and high visual features is presented. These features have been obtained by exploiting concepts from and art theory (Itten, 1973; Valdez and Mehrabian, 1994), or exploited in image retrieval (Stottinger et al., 2009) and image classification (Datta et al., 2006; Wei-ning et al., 2006) tasks. They selected 17 visual features, categorized in 4 groups: *colour* (e.g., mean saturation and brightness, hue statistics, colourfulness measure, Itten contrast (Itten, 1973), colour histogram, etc.), *texture* (e.g., wavelet textures, correlation, contrast, homogeneity, and energy for the HSB channels, etc.), *composition* (number of segments after the waterfall segmentation, depth of field (DOF), rule of thirds, etc.) and *content* (number of faces, pixels of the biggest face, number of skin pixels, etc.).

Most of the mentioned works combine huge number of hand-crafted visual features. Although all the exploited features have been proven to have a direct influence on the perceived emotion by previous studies, there is not agreement about which of them give the most of the contribution on the aimed task. Besides the selection of proper hand-crafted features, designed with the aim to encode the sentiment content conveyed by images, there are other kind of approaches that lean on representation learning techniques based on Deep Learning (Chen et al., 2014; Xu et al., 2014; You et al., 2015). By employing such representation methods, image features are learned from the data. This avoid the designing of a proper set of feature, because the system automatically learns how to extract the needed information from the input data. These methods requires huge amounts of labelled training data, and an intensive learning phase, but obtain better performances in general. Another approach, borrowed from the image retrieval methods, consists on combining textual and visual information through multimodal embedding systems (Katsurai and Satoh, 2016; Ortis et al., 2018). In this case, features taken from different modalities (e.g., visual, textual, etc.) are combined to create a common vector space in which the correlations between projections of the different modalities are maximized (i.e., an embedding space).

4 PROBLEM PRESENTATION

In this section we discuss the key components that affect the sentiment. This allows to better focus the related sub-issues which form the Image Sentiment Analysis problem and support the designing of more robust approaches. An opinion consists of two main components: a target (or topic), and a sentiment. The opinions can be taken from more than one person, and opinions can change over time. This means that the system has to take into account also the opinion holder and the time an opinion is expressed. In the case of Sentiment Analysis on visual contents there are some differences. When the input is a text, Sentiment Analysis can exploit NLP (Natural Language Processing) techniques to extract the different part of a sentence, and associate each word with a specific meaning. When the input is an image the task to associate visual features to sentiment categories or polarity scores results challenging. In the following paragraphs each of the sentiment components previously mentioned (i.e., entity, aspect, holder and time) are discussed in the context of Image Sentiment Analysis.

4.1 Sentiment Entity and Aspects

The entity is the subject (or target) of the analysis. In the case of Image Sentiment Analysis the entity is the input image. In general, an entity can be viewed as a set of "parts" and "attributes". The set of the entity's parts, its attributes, plus the special aspect "GENERAL" forms the set of the "aspects". This structure can be transferred to the image domain considering different levels of visual features. The parts of an image can be defined by considering a set of sub-images. This set can be obtained by exploiting several Computer Vision techniques, such as background/foreground extraction, image segmentation, multi object recognition or dense captioning (Karpathy and Fei-Fei, 2015). The attributes of an image regards its aesthetic quality features, often obtained by extracting low-level features. Exploiting this structured image hierarchy, a sentiment score can be inferred for each aspect. Finally, the partial scores can be properly combined to obtain the sentiment classification (e.g., data can be used as input features of a regression model). One can consider the concept associated to the image context. For this purpose, several works about personal contexts (Ortis et al., 2017; Furnari et al., 2018) and scene recognition can be exploited from the visual view, and the inferred concepts can be used to extract the associated sentiment. Moreover, sentiment scores can be further extracted from image parts and attributes. Instead of representing the image parts as a set of sub-images, an alternative approach can rely on a textual description of the depicted scene. The description of a photo can be focused on a specific task of image understanding. By changing the task, we can obtain multiple descriptions of the same image from different points of view. Then, these complementary concepts can be combined to obtain the above described structure. Most of the existing works in analysing social media exploit textual information manually associated to images by performing textual Sentiment Analysis. Although the text associated to social images is widely exploited in the state-of-the-art to improve the semantics inferred from images, it can be a very noisy source because it is provided by the users. Starting from this observation, the authors of (Ortis et al., 2018) presented a work on Image Polarity Prediction exploiting Objective Text extracted directly from images, and experimentally compared such text with respect to the Subjective (i.e., user provided) text information commonly used in the state-of-the-art approaches. Such approach provides an alternative user-independent source of text which describes the semantic of images, useful to address the issues related to the inherent subjectivity of the text associated to images.

4.2 Sentiment Holder

Almost all the works in Image Sentiment Analysis ignore the sentiment holder, or implicitly consider only the sentiment of the image publisher. In this context at least two holders can be defined: the image owner and the image viewer. Considering the example of an advertising campaign, understanding the connections between the sentiment intended by the owner and the actual sentiment induced to the viewers is crucial.

These days, the social media platforms provide a very powerful mean to retrieve real-time and large scale information of people reactions toward topics, events and advertising campaigns. This branch of research can be useful to understand the relation between the affect concepts of the image owner and the evoked viewer ones, allowing new user centric applications. User profiling helps personalization, which is very important in the field of recommendation systems.

4.3 Time

Although almost all the aforementioned works ignore this aspect, the emotion evoked by an image can change depending on the time. This sentiment component can be ignored the most of times, but in specific cases is determinant. For example, the sentiment evoked by an image depicting the World Trade Center is presumably different if the image is shown before or after 9/11.

Although there are not works on Image Sentiment Analysis that analyse the changes of image sentiments over time, due to the specificity of the task and the lack of image datasets, there are several works that exploits the analysis of images over time focused on specific cognitive and psychology applications. As an example, the work in (Reece and Danforth, 2017) employed a statistical framework to detect depression by analysing the sequence of photos posted on Instagram. The findings of this paper suggest the idea that variations in individual psychology reflect in the use social media by the users, hence they can be computationally detected by the analysis of the user's posting history. In (Khosla et al., 2012) the authors studied which objects and regions of an image are positively or negatively correlated with memorability, allowing to create memorability maps for each image.

5 CURRENT CHALLENGES

So far we discussed on the current state of the art in Image Sentiment Analysis, describing the related issues, the different employed approaches and features. This section aims to introduce some additional challenges and techniques that can be investigated.

5.1 Image Popularity Prediction

One of the most common application field of Image Sentiment Analysis is related to social marketing campaigns. In the context of social media communication, several companies are interested to analyse the level of people engagement with respect to social posts related to their products. This can be measured as the number of post's views, likes, shares or by the analysis of the comments.

The level of engagement of an image posted on a social network is usually referred as "Image Popularity". So far, researches have been trying to gain insights into what features make an image popular.

In (Khosla et al., 2014) the authors proposed a log-normalized popularity score that has been then commonly used in the community. Let c_i be a measure of the engagement achieved by a social media item (e.g., number of likes, number of views, number of shares, etc.), also known as popularity measure. The popularity score of the i^{th} item is defined as follows:

$$score_i = \log\left(\frac{c_i}{T_i} + 1\right) \quad (1)$$

where T_i is the number of days since the uploading of the image on the Social Platform.

Although the task of image popularity prediction is rather new, there are interesting datasets available for the development of massive learning systems (i.e., deep neural networks). The Micro-Blog Images 1 Million (MBI-1M) dataset is a collection of 1M images from Twitter, along with accompanying tweets and metadata. The dataset was introduced by the work in (Cappallo et al., 2015b). The MIR-1M dataset (Huiskes et al., 2010) is a collection of 1M photos from Flickr. These images have been selected considering the interestingness score used by Flickr to rank images. The Social Media Prediction (SMP) dataset is a large-scale collection of social posts, recently collected for the ACM Multimedia 2017 SMP Challenge [3]. This dataset consists of over 850K posts and 80K users, including photos from VSO (Borth et al., 2013) as well as photos collected from personal users' albums (Wu et al., 2016).

Popularity Dynamics: Equation 1 normalizes the number of interactions reached by an image by dividing the engagement measure by the time. However, the measures c_i related to social posts are cumulative values as they continuously collect the interactions between users and the social posts during their time on-line. Therefore, this normalization will penalize social media contents published in the past with respect to more recent contents, especially when the difference between the dates of posting is high. Moreover, the most of the engagement obtained by a social media item is achieved in the first week (Valafar et al., 2009), then the engagement measures become more stable. Therefore, it would be interesting to develop methods able to predict the popularity score evolution over time, instead of just estimating a normalized score.

5.2 Image Virality Prediction

A recent emerging task, closely related to image popularity, is the prediction of the level of virality of an image. The image virality is defined as the quality of a visual content (i.e., images or videos) to be rapidly and widely spread on social networks (Alameda-Pineda et al., 2017). Differently than popularity, the virality score takes into account also the number of resubmission of an image by different users. Therefore, images that became popular when they are posted, but not reposted, are not considered to be viral (Deza and Parikh, 2015). The work in (Alameda-Pineda et al., 2017) focused on understanding the influence of image parts on its virality. In particular, the authors presented a method able to perform the prediction of the virality score and the localicazion of areas in images that are responsable for making the image viral.

[3] Challenge webpage: https://social-media-prediction.github.io/MM17PredictionChallenge

5.3 Relative Attributes

Several Image Sentiment Analysis works aim to associate an image one sentiment label over a set of emotional categories or attributes. However, given a set of images that have been assigned to the same emotional category (e.g., joy), it would be interesting to determine their ranking with respect the specific attribute. Such a technique could suggest, for example, if a given image *A* conveys more "joy" than another image *B*. For this purpose, several works on relative attributes can be considered (Parikh and Grauman, 2011; Altwaijry and Belongie, 2013; Fan et al., 2013; Yu and Grauman, 2015).

5.4 Sentiment and Ideograms

The emoticons are textual shorthand that have been introduced to allow the writer to express feelings and emotions with respect to a textual message. It helps to express the correct intent of a text sentence, improving the understanding of the message. The emoticons are used to emulate visual cues in textual communications with the aim to express or explicitly clarify the writer's sentiment. Indeed, in real conversations the sentiment can be inferred from visual cues such as facial expressions, pose and gestures. However, in textual based conversations, the visual cues are not present. The authors of (Hogenboom et al., 2013) tried to understand if emoticons could be useful as well on the textual Sentiment Analysis task. In particular, they investigated the role that emoticons play in conveying sentiment and how they can be exploited in the field of Sentiment Analysis. A step further the emoticon, is represented by the emoji. An emoji is an ideogram representing concepts such as weather, celebration, food, animals, emotions, feelings, and activities, besides a large set of facial expressions. They have been developed with the aim to allow more expressive messages. Emojis have become extremely popular in social media platforms and instant messaging systems. In (Cappallo et al., 2015a), the authors exploited the expressiveness carried by emoji, to develop a system able to generate an image content description in terms of a set of emoji. The focus of this system is to use emoji as a means for image retrieval and exploration. Indeed, it allows to perform an image search by means of a emoji-based query. This approach exploits the expressiveness conveyed by emoji, by leaning on the textual description of these ideograms. The work in (Cappallo et al., 2018) studied the ways in which emoji can be related to other common modalities such as text and images, in the context of multimedia research. This work also presents a new dataset that contains examples of both text-emoji and image-emoji relationships. Most of them contains also strong sentiment properties. In (Kralj Novak et al., 2015) the authors presented a sentiment emoji lexicon named Emoji Sentiment Ranking. In this paper, the sentiment properties of the emojis have been deeply analyzed (considering text written in 13 different languages), and some interesting conclusions have been highlighted. For each emoji, the *Emoji Sentiment Ranking* provides its associated positive, negative and neutral scores. The authors found that the sentiment scores and ranking associated to emojis remain stable among different languages. This property is very useful to overcome the difficulties addressed in multilingual contexts. The results and the insights obtained in the mentioned works could be combined to exploit the sentiment conveyed by emoji on the task of Image Sentiment Analysis.

6 CONCLUSIONS AND FUTURE WORKS

The proposed paper provides a review of relevant publications on Image Sentiment Analysis, and presents an overview of the state of the art in the field. Principles of design of Image Sentiment Analysis systems are presented and discussed under three main points of view: emotional models, dataset definition and feature design. The components that can affect the sentiment toward an image in different ways are defined and analysed. A description of new challenges is also presented. There is a wide range of research on identification of basic emotions, but the 24 emotions model defined in Plutchnik's theory (Plutchik, 1980) represents a well established and used model. The Machine Learning techniques and the recent Deep Learning methods are able to obtain impressive results as long as these systems are trained with very large scale datasets (e.g., VSO (Borth et al., 2013)). Such datasets can be easily obtained by exploiting the social network platforms by which people share their pictures every day. These datasets allowed the extensive use of Machine Learning systems that requires large scale amount of data to converge. So far, there is not an established strategy to select of visual features that allows to address the problem. Most of the previous exploited features demonstrated to be useful, but recent results on Image Sentiment Analysis suggest that it's worth investigating the use of representation learning approaches such as Convolutional Neural Networks and multi-modal embedding.

In future works the presented study will be extended, considering publications that address specific aspects of the problems mentioned in the paper (e.g., features, models, metrics, dataset definition, etc.), as well as scientific works that tackle specific tasks including the challenges that have been briefly presented in the proposed paper (e.g., popularity, relative attributes, virality, etc.). Furthermore, proper cases of study aimed to highlight the concepts described in the paper will be prepared and evaluated, with the aim to support the key aspects of each addressed issue.

REFERENCES

Alameda-Pineda, X., Pilzer, A., Xu, D., Sebe, N., and Ricci, E. (2017). Viraliency: Pooling local virality. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 484–492.

Altwaijry, H. and Belongie, S. (2013). Relative ranking of facial attractiveness. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 117–124.

Battiato, S., Farinella, G. M., Milotta, F. L., Ortis, A., Addesso, L., Casella, A., D'Amico, V., and Torrisi, G. (2016). The social picture. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 397–400. ACM.

Borth, D., Ji, R., Chen, T., Breuel, T., and Chang, S.-F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232. ACM.

Bradley, M. M. (1994). Emotional memory: A dimensional analysis. *Emotions: Essays on emotion theory*, pages 97–134.

Campos, V., Jou, B., and i Nieto, X. G. (2017). From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction. *Image and Vision Computing*, 65:15 – 22. Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.

Campos, V., Salvador, A., Giró-i Nieto, X., and Jou, B. (2015). Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction. In *Proceedings of the 1st International Workshop on Affect &#38; Sentiment in Multimedia*, ASM '15, pages 57–62, New York, NY, USA. ACM.

Cappallo, S., Mensink, T., and Snoek, C. G. (2015a). Image2emoji: Zero-shot emoji prediction for visual media. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1311–1314. ACM.

Cappallo, S., Mensink, T., and Snoek, C. G. (2015b). Latent factors of visual popularity prediction. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 195–202. ACM.

Cappallo, S., Svetlichnaya, S., Garrigues, P., Mensink, T., and Snoek, C. G. M. (2018). The new modality: Emoji challenges in prediction, anticipation, and retrieval. *IEEE Transactions on Multimedia*. Pending minor revision.

Chen, T., Borth, D., Darrell, T., and Chang, S.-F. (2014). Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*.

Colombo, C., Del Bimbo, A., and Pala, P. (1999). Semantics in visual information retrieval. *IEEE Multimedia*, 6(3):38–53.

Dan-Glauser, E. S. and Scherer, K. R. (2011). The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance. *Behavior research methods*, 43(2):468–477.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, pages 288–301. Springer.

Deza, A. and Parikh, D. (2015). Understanding image virality. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1818–1826.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P., Scherer, K., Tomita, M., and Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer.

Fan, Q., Gabbur, P., and Pankanti, S. (2013). Relative attributes for large-scale abandoned object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2736–2743.

Furnari, A., Battiato, S., and Farinella, G. M. (2018). Personal-location-based temporal segmentation of egocentric video for lifelogging applications. *Journal of Visual Communication and Image Representation*, 52:1–12.

Hayashi, T. and Hagiwara, M. (1998). Image query by impression words-the iqi system. *IEEE Transactions on Consumer Electronics*, 44(2):347–352.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., and Kaymak, U. (2013). Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 703–710. ACM.

Huiskes, M. J., Thomee, B., and Lew, M. S. (2010). New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. In *Proceedings of the international conference on Multimedia information retrieval*, pages 527–536. ACM.

Itten, J. (1973). *The Art of Color: The Subjective Experience and Objective Rationale of Color.* John Wiley & Sons Inc.

Jufeng Yang, Dongyu She, M. S. (2017). Joint image emotion classification and distribution learning via deep convolutional neural network. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3266–3272.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Katsurai, M. and Satoh, S. (2016). Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2837–2841. IEEE.

Khosla, A., Das Sarma, A., and Hamid, R. (2014). What makes an image popular? In *Proceedings of the 23rd international conference on World wide web*, pages 867–876. ACM.

Khosla, A., Xiao, J., Torralba, A., and Oliva, A. (2012). Memorability of image regions. In *Advances in Neural Information Processing Systems*, pages 305–313.

Kralj Novak, P., Smailović, J., Sluban, B., and Mozetič, I. (2015). Sentiment of emojis. *PLOS ONE*, 10(12):1–22.

Lang, P. J. (1993). The network model of emotion: Motivational connections. *Perspectives on anger and emotion: Advances in social cognition*, 6:109–133.

Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (1999). International affective picture system (iaps): Technical manual and affective ratings. *Gainesville, FL: The Center for Research in Psychophysiology, University of Florida*.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Machajdik, J. and Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92. ACM.

Mikels, J. A., Fredrickson, B. L., Larkin, G. R., Lindberg, C. M., Maglio, S. J., and Reuter-Lorenz, P. A. (2005). Emotional category data on images from the international affective picture system. *Behavior research methods*, 37(4):626–630.

Ortis, A., Farinella, G. M., D'Amico, V., Addesso, L., Torrisi, G., and Battiato, S. (2017). Organizing egocentric videos of daily living activities. *Pattern Recognition*, 72(Supplement C):207 – 218.

Ortis, A., Farinella, G. M., Torrisi, G., and Battiato, S. (2018). Visual sentiment analysis based on on objective text description of images. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological bulletin*, 49(3):197.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Parikh, D. and Grauman, K. (2011). Relative attributes. In *IEEE International Conference on Computer Vision*, pages 503–510. IEEE.

Peng, K.-C., Chen, T., Sadovnik, A., and Gallagher, A. C. (2015). A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 860–868.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31.

Reece, A. G. and Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1):15.

Russell, J. A. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Schmidt, S. and Stock, W. G. (2009). Collective indexing of emotions in images. a study in emotional information retrieval. *Journal of the American Society for Information Science and Technology*, 60(5):863–876.

Shaver, P., Schwartz, J., Kirson, D., and O'connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061.

Siersdorfer, S., Minack, E., Deng, F., and Hare, J. (2010). Analyzing and predicting sentiment of images on the social web. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 715–718. ACM.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Stottinger, J., Banova, J., Ponitz, T., Sebe, N., and Hanbury, A. (2009). Translating journalists' requirements into features for image search. In *15th International Conference on Virtual Systems and Multimedia.*, pages 149–153. IEEE.

Sun, M., Yang, J., Wang, K., and Shen, H. (2016). Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. In *IEEE International Conference on Multimedia and Expo.*, pages 1–6. IEEE.

Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment.

Vadicamo, L., Carrara, F., Cimino, A., Cresci, S., Dell'Orletta, F., Falchi, F., and Tesconi, M. (2017). Cross-media learning for image sentiment analysis in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 308–317.

Valafar, M., Rejaie, R., and Willinger, W. (2009). Beyond friendship graphs: a study of user interactions in flickr. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 25–30. ACM.

Valdez, P. and Mehrabian, A. (1994). Effects of color on emotions. *Journal of experimental psychology: General*, 123(4):394.

Wei-ning, W., Ying-lin, Y., and Sheng-ming, J. (2006). Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3534–3539. IEEE.

Wu, B., Cheng, W.-H., Zhang, Y., and Mei, T. (2016). Time matters: Multi-scale temporalization of social media popularity. In *Proceedings of the 2016 ACM on Multimedia Conference (ACM MM)*.

Xu, C., Cetintas, S., Lee, K.-C., and Li, L.-J. (2014). Visual sentiment prediction with deep convolutional neural networks. *arXiv preprint arXiv:1411.5731*.

Yang, J., Sun, M., and Sun, X. (2017). Learning visual sentiment distributions via augmented conditional probability neural network. In *AAAI*, pages 224–230.

Yanulevskaya, V., Van Gemert, J., Roth, K., Herbold, A.-K., Sebe, N., and Geusebroek, J.-M. (2008). Emotional valence categorization using holistic image features. In *15th IEEE International Conference on Image Processing*, pages 101–104. IEEE.

You, Q., Luo, J., Jin, H., and Yang, J. (2015). Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 381–388. AAAI Press.

Yu, A. and Grauman, K. (2015). Just noticeable differences in visual attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2416–2424.

Yuan, J., Mcdonough, S., You, Q., and Luo, J. (2013). Sentribute: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 10. ACM.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495.