Automatic Emotional Text Annotation Using Facial Expression Analysis

Alex Mircoli
Supervisors: Alessandro Cucchiarelli, Claudia Diamantini, Domenico Potena

Department of Information Engineering
Università Politecnica delle Marche, Ancona, Italy
a.mircoli@pm.univpm.it

Abstract. Sentiment analysis calls for large amounts of annotated data, which are usually in short supply and require great efforts in terms of manual annotation. Furthermore, the analysis is often limited to text polarity and writer's emotions are ignored, even if they provide valuable information about writer's feelings that might support a large number of applications, such as open innovation processes. Our research is hence aimed at developing a methodology for the automatic annotation of texts with regard to their emotional aspects, exploiting the correlation between speech and facial expressions in videos. In the present work we describe the main ideas of our proposal, presenting a four-phase methodology and discussing the main issues related to the selection of input frames and the processing of emotions resulting from facial expressions analysis.

Keywords: Sentiment Analysis · Automated text annotation · Facial expression analysis · Emotion detection

1 Introduction

In recent years, the rapid growth of social networks, personal blogs and review sites has made available an enormous amount of user-generated content. Such data often contain people's opinions and emotions about a certain topic; that information is considered authentic, as in the above contexts people usually feel free to express their thoughts. Therefore, the analysis of this user-generated content provides valuable information on how a certain topic or product is perceived by users, allowing firms to address typical marketing problems as, for instance, the evaluation of customer satisfaction or the measurement of the appreciation of a new marketing campaign. Moreover, the analysis of customers' opinions about a certain product helps business owners to find out possible issues and may suggest new interesting features, thus representing a valid tool for open innovation.

For this reason, in the last years many researchers (e.g., [1], [2]) focused on techniques for the automatic analysis of writer's opinion, generally referred to as *sentiment analysis* [3]. In its most common meaning, the term refers to the analysis of the text polarity, that is the evaluation of positiveness or negativeness of the author's

X. Franch, J. Ralyté, R. Matulevičius, C. Salinesi, and R. Wieringa (Eds.):
CAiSE 2017 Forum and Doctoral Consortium Papers, pp. 188-197, 2017.

view towards a particular entity. Due to the intrinsic complexity of the human language, this task offers several challenges, some of which, namely the automatic detection of the scope of negation [4] and the disambiguation of polysemous words [5], have been addressed in our previous work.

A more general definition of sentiment analysis also involves the computational treatment of the emotional aspects of text [6]. In this paper we adopt this broader definition and, in particular, we focus on the detection of feelings and emotions in texts.

Sentiment analysis is generally performed using two different approaches, that respectively rely on annotated lexical resources (lexicon-based techniques) [7] and deep neural networks (learning-based techniques) [8]. Learning-based techniques usually reach high accuracy but they need considerable amounts of annotated training data. Moreover, they are very domain-specific, so the creation of a new annotated dataset is required whenever the model needs to be retrained for a different domain. On the other hand, the lexicon-based approach requires the availability of large corpora. At the moment, few open access corpora exist (e.g., SentiWordNet [9]) and they are only available for English. Furthermore, emotions are considered by few corpora (e.g., [10]) and typically for small sets of words. In conclusion, both approaches rely on the existence of large amounts of annotated data, which require great efforts in terms of manual annotation. Therefore, there is a need for techniques to automatically (and objectively) annotate emotions in text, in order to dynamically create language- and domain-specific corpora for sentiment analysis.

To this purpose, our research aims at developing a novel methodology for the automatic creation of emotionally annotated corpora through the analysis of speakers' facial expressions in subtitled videos. These corpora will enable the analysis of the emotional aspects of user-generated content, in order to provide valuable insights about, for instance, consumer perception or voters' opinion. We start from the following research questions (RQ):

- RQ1: What is the state of the art in sentiment analysis?
- RQ2: How can we automatically annotate a corpus w.r.t emotions expressed in text?
- RQ3: How can the resulting emotional annotation be evaluated?
- RQ4: How can we exploit our system to enhance traditional Business Intelligence applications?

Due to the popularity of video-sharing platforms, such as YouTube, a multitude of subtitled videos has been made publicly available: as a consequence, an automatic annotation technique allows for the analysis of large amounts of text data. The choice of analyzing facial expressions is driven by the consideration that they are unambiguous non-verbal clues that can be exploited to precisely estimate the speaker's emotional state. Furthermore, in [11] is empirically verified that there is a strong correlation between facial expressions and speech, so that spontaneous speakers are expected to exhibit emotions consistently with their speech.

In this work, we model human emotions according to Ekman's theory [12], that states the existence of six archetypal emotions (i.e., anger, disgust, fear, happiness, sadness, surprise). Text annotation by means of the analysis of these basic facial expressions shows several advantages: (i) every emotion can be viewed as a combination of the six basic emotions and hence annotating with respect to basic emotions allows for the representation of every human emotion; (ii) the technique is language- and domain-independent, so emotional annotations can be carried out for every language and scope of application, and finally (iii) facial expressions are demonstrated to be universal [13], that is unrelated to speaker's personal characteristics (e.g., sex, ethnicity).

The rest of the paper is structured as follows: next Section is devoted to present some relevant related works on sentiment analysis and automatic annotation of corpora. The methodology for the automatic text annotation is proposed in Section 3, along with the description of the current status of the research. Finally, Section 4 draws conclusions and discusses future directions of research.

2 Related Work

In recent literature, several different approaches have been proposed for the analysis of writer's opinions. Socher et al. [16] present a recursive neural tensor network that reaches a state-of-the-art classification accuracy when trained on annotated parse trees. The main disadvantage of the technique is that it requires a huge amount of training data: more than 600,000 human annotations were needed to train the original model and training the network on a different domain would require similar efforts. Go et al. [17] propose a solution based on distant supervision, in which training data consists of tweets with emoticons, that reaches an accuracy similar to [16] without requiring manual annotations. However, both approaches only analyze the text polarity, without considering the emotional content of text.

The analysis of writer's emotions is performed in [18], where authors consider emotion-word hashtags as labels for emotions in tweets. Nevertheless, their approach differs from ours since they are more focused on the analysis of personality traits and they do not consider para-verbal indicators. To the best of our knowledge, the analysis of para- and non-verbal communication, such as facial expressions, for the purpose of automatic emotional text annotation has received no attention in recent literature: although Busso et al. [19] propose to analyze facial expressions and speech tone in order to detect emotions, they do not exploit this information to build annotated corpora.

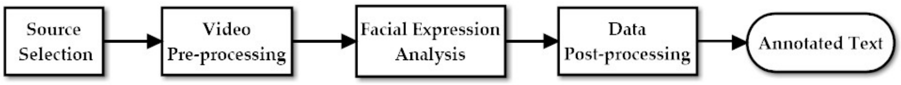

Fig. 1. The methodology for the automatic text annotation through facial expression analysis

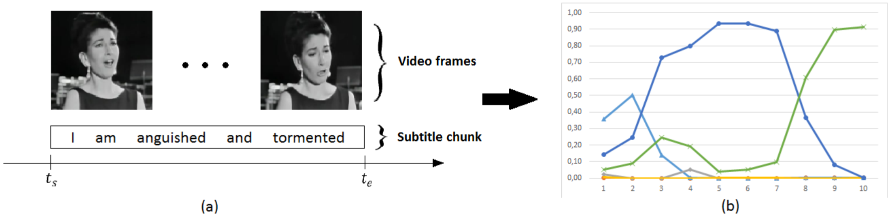

Fig. 2. Example: extraction of emotion vectors from a video

3 Automatic Text Annotation

In this Section we describe the methodology for the automatic emotional text annotation based on the analysis of speakers' facial expressions in videos. As depicted in Fig.1, the proposed methodology consists of four steps: first, a data source is selected, with particular attention to some issues that could preclude the feasibility and/or the accuracy of emotion detection. Afterwards, in the second step frames are extracted and analyzed using a face recognition software, in order to filter out scenes with zero or multiple faces. The facial expressions of people appearing in the remaining frames are then analyzed (step three) and the resulting emotion vectors are assigned to the corresponding subtitle chunk. Finally, in step four the emotion associated to each subtitle chunk is computed starting from the emotion vectors.

An example of this procedure is depicted in Fig.2, where frames belonging to the subtitle chunk "I am anguished and tormented" (Fig. 2(a)) are processed by a facial expression analyzer, whose output is a set of emotion vectors, that are plotted in Fig.2(b). In the line chart, the sadness expression (blue line, circular marker) dominates the central part of the graph (frames 3-7), while disgust expressions (green line, x-shaped marker) appear on the final frames.

A more detailed discussion of each methodology step is presented in the following subsections, along with the description of the current state of the research.

3.1 Source Selection

The first step involves the selection of the data source. This is a crucial phase, as the quality of the final annotated corpus strongly depends on the selection of a suitable set of input videos. Even if emotions are expressed through universally shared patterns, there are several factors that can impact on speaker's spontaneity and expressiveness and that must be considered in selecting the video categories to be analyzed. A list of issues that we faced in our preliminary video scouting activity, and can potentially impact on the quality of text annotation, includes:

- *lack of expressiveness*: in some video typologies, such as news reports or product reviews, speakers are required to maintain an expressionless face, in order to give objectivity to their speech. As a consequence, the analysis of their facial expressions can be misleading, as emotion-bearing words could be annotated as neutral.

- *interpretation*: in case of movies or theatrical monologues, some actors are required to play characters having a specific personality. In such circumstances, facial expressions are altered by acting: a criminal, for instance, might have a scary face even when talking about happy things.
- *reported speech*: when people report what another person has said, facial expressions reflect their personal feelings and hence detected emotions can be in contrast with the original meaning of the sentences.
- *external factors*: external factors impacting on speaker's mood can affect the correlation between speech and facial expressions. For instance, an eyewitness interviewed immediately after a plane crash would probably show fear expressions, regardless of the specific words he is pronouncing.
- *subtitles quality*: video and subtitles might not be in synchronization, thus facial expressions could not correspond to subtitle text, or subtitles may not be accurate, as they have been automatically generated.

The above-listed issues are noise factors that cannot be totally avoided but the selection of a proper data source can effectively impact on their presence in analyzed videos. In this preliminary phase we limited to explore several different data sources and manually select those that we considered more suitable to the purpose of our analysis.

Nevertheless, some of these problems can be automatically detected in videos: for example, interviews and news reports can often be identified through the analysis of the video title, while in some video sharing sites (e.g., YouTube) there are dedicated APIs to find out if subtitles have been automatically generated.

3.2 Video Pre-processing

After selecting a proper data source, each video is subject to a preprocessing phase, whose output is the set of frames F to be analyzed. From a computational perspective, a desirable property for F is that $|F| \ll |V|$, where V is the entire set of video frames, since the analysis of facial expressions is a computationally-intensive task, that can take up to five seconds for each frame on state-of-the-art facial analysis tools. Considering that typical frame rates are around 30 fps, the analysis of every frame of a 5-minute video may require up to 150 minutes to be performed, making infeasible large-scale annotations. Anyway, many frames may be discarded without significant loss of information: for instance, since the purpose of the methodology is the annotation of subtitles on the basis of the concomitant speaker's expressions, frames not related to any subtitle chunk may also be discarded. Apart from this preliminary operation, video pre-processing consists of sampling and filtering.

Sampling. The choice of the sampling rate (sr), that is the number of fps to be extracted, has implications on both speed and accuracy. A high value for the parameter

sr implies a higher number of frames to be analyzed, with a consequent increase in the execution time of the facial expressions analysis. Moreover, a speaker is expected to exhibit almost identical expressions in a block of consecutive frames, then the analysis of the entire block is redundant.

On the other hand, by choosing a small value for sr (e.g., $sr < 1$) there is the risk to extract many irrelevant frames, such as those where there are transitions from an expression to another. Furthermore, facial expressions are somewhat dependent on the concomitant phonatory movements. For instance, the pronunciation of the vowel [ɑ] requires speakers to widely open their mouth, that could be interpreted as a surprise expression. As a consequence, the analysis of a too small amount of frames is error-prone, especially in presence of speakers with a great articulation of open vowels.

We performed some preliminary experiments in order to find a value for sr that would balance speed and accuracy. We found that $2 \leq sr \leq 4$ offers a classification accuracy comparable to the analysis of every frame, while reducing of approximately one order of magnitude the whole execution time.

Filtering. Some sampled frames should be discarded as they do not contain useful information:

- frames with zero or multiple faces, as well as frames containing speakers not facing the camera, cannot be analyzed due to the lack (or the excess, with the consequent problem of identifying speaker's face) of correctly recognizable faces.
- in case of two temporally close subtitle chunks (e.g., when they both belong to the same sentence) s_j and s_{j+1}, the speaker's facial expressions in the first frames of s_{j+1} could be related to s_j, because emotions may remain on speaker's face up to 2-2.5 seconds after the end of the related sentence [11]. Therefore, if the ending time $t_j^{(e)}$ of s_j is close or coincident to the starting time $t_{j+1}^{(s)}$ of s_{j+1}, the first k frames related to s_{j+1} may be excluded from F, where the parameter k has to be chosen empirically.

The filtering step can be automated through a face analysis tool: in particular, in our system we use Microsoft Face API[1], that provides information about the number and the pose (i.e., facing/not facing camera) of faces in an image.

3.3 Facial Expression Analysis

As a result of the video pre-processing phase, each s_j is associated with a set of frames $F_j \subseteq F$, where $F_j = \{f_{\lceil t_j^{(s)} \cdot sr \rceil}, f_{\lceil t_j^{(s)} \cdot sr \rceil + 1}, \ldots, f_{\lceil t_j^{(e)} \cdot sr \rceil}\}$; the symbol $\lceil \cdot \rceil$ denotes the $ceil(\cdot)$ operator (alternatively, the $floor(\cdot)$ operator may be used). In this step, for each frame $f_i \in F_j$ we analyze the speaker's facial expression with respect to Ekman's theory of six basic emotions (i.e., anger, disgust, fear, happiness, sadness, surprise) and we obtain the emotion vector $e_i \epsilon E$, defined as

[1] https://www.microsoft.com/cognitive-services/en-us/emotion-api

$e_i = [e_i^{(an)} e_i^{(di)} e_i^{(fe)} e_i^{(ha)} e_i^{(sa)} e_i^{(su)}]^T$, where E is the emotion matrix of the video and $e_i^{(\alpha)}$ represents the value of the α emotion in f_i. At the end of the analysis, each subtitle chunk s_j is associated with the emotion matrix $E_{F_j} = E_{\lceil t_j^{(s)} \cdot sr \rceil : \lceil t_j^{(e)} \cdot sr \rceil}$, where $E_{i:j}$ denotes the $(j - i + 1)$ columns of E starting from the i-th column.

At the moment, we perform the facial expression analysis through the free version of Microsoft Emotion API. The facial expression analysis is performed with respect to eight classes: in addition to Ekman's basic expressions, the software evaluates the contempt and the neutral expressions. The value of the latter is calculated as the 1's complement of the sum of the other vector components, each of which has a value in $[0,1]$.

In early experiments we found that the analysis of facial expressions of speaking people is a challenging task, since the degree of mouth opening (that is considered a key feature by the facial expression analyzer) of a speaker strongly depends on the articulation of the speech. Although we limited our experiments to Microsoft Emotion API, it is plausible that other tools may have the same issue, since many facial expression analysis techniques in literature relies on this feature (e.g., [20]). We hypothesized that mouths could be removed from images without compromising the analysis, as the software may still rely on the position of eyes, forehead and eyebrows as emotion markers. We tested our hypothesis on a small test set and we noticed a 20% increase in classification accuracy when mouths were manually hidden before performing the facial expression analysis.

3.4 Data Post-processing

The outputs of the previous step are the emotion matrix E and a set of annotations in the form $s_j \rightarrow E_{F_j}$. This level of annotation provides information about the distribution of emotions in each frame, while we are more interested in a text-level annotation. The final form of the annotated text strictly depends on the kind of application it is intended for: training data for machine learning are usually in the form $s_j \rightarrow c_j$, where the class $c_j \in C$ corresponds to a basic emotion, while annotated corpora for lexicon-based sentiment analysis may also be in the form $s_j \rightarrow \hat{e}_j$, where $\hat{e}_j$ is a vector containing an aggregate value for each emotion.

In general, the first step of the post-processing phase is the definition of the aggregation function f:

$$f\colon E_j \rightarrow \hat{e}_j \tag{1}$$

that applies an aggregation operator (e.g., max, avg) to every row of E_j, outputting a vector $\hat{e}_j$ having a single aggregate value for each emotion.

To bind each subtitle chunk to a class label c_j, the transform function g is applied:

$$g\colon \hat{e}_j \rightarrow c_j \tag{2}$$

in order to evaluate the dominant emotion and assign the related class label to s_j. The choice of g is non-trivial, as in some situations there is no correspondence between the maximum value of $\hat{e}_j$ and the actual dominant emotion. For instance, in case of noisy scenes, the facial expression analyzer would probably assign lower values to emotions, as it is not able to detect enough emotion markers in speaker's face. In this scenario, the function g should discard the highest value in $\hat{e}_j$, whenever it is related to the neutral emotion and there is another emotion with a sufficiently high value.

At the moment, in our implementation we set $f = avg(\cdot)$ and $g = \max(\cdot)$ but other solutions are under investigation: for example, we are interested in performing experiments using $f = max(\cdot)$. In preliminary experiments we found that the annotation accuracy of the system depends on the quality of input videos (e.g., expressive speakers, synchronized subtitles). For some videos we reached a remarkable 66% classification accuracy in 8-class emotional annotation. A major advantage of our technique emerged during the analysis of our results: when a sentence is automatically annotated starting from the sentiment of each word, it is not always possible to correctly capture its overall sentiment, which instead emerges from speaker's expression. For instance, the phrase "throw myself in the Arno" is correctly annotated as "sadness", although in the phrase there are not words related to sadness.

In our preliminary analysis, we also noticed that some words frequently co-occur in sentences related to a specific emotion, suggesting that there is a correlation between these n-grams and the emotion. For this reason, we plan to perform a large-scale analysis to further investigate this phenomenon.

4 Future Directions

Our work is organized in terms of a three years Ph.D. research project. The first year was dedicated to the definition of the methodology, in addition to an early experimentation in order to validate the research idea. In the next two years, we plan to move along two main directions: firstly, the study of the best f and g functions for data post-processing, in particular to deal with noisy scenes, where emotion vectors show wide variations even among frames containing similar facial expressions. An appropriate choice of f and g will reduce misclassifications, thus allowing for a more accurate detection of people's emotions in texts.

Secondly, the extension of experiments to larger datasets, in order to better evaluate the classification accuracy and perform statistical analysis to find correlations between specific couples <n-gram, emotion>, with the purpose of building an emotionally annotated corpus for sentiment analysis.

Moreover, we intend to further investigate the effectiveness of hiding mouths and possibly automate this procedure.

Finally, since the quality of the annotation strongly depends on the accuracy of the facial expression analysis, we are interested in evaluating different face analysis tools, in order to compare their performance and choose the best solution.

References

1. Chien, C. C., You-De, T.: Quality evaluation of product reviews using an information quality framework. Decision Support Systems 2011, Vol. 50, pp.755–68 (2011)
2. Lai, S., Xu, L., Zhao, J.: Recurrent Convolutional Neural Networks for Text Classification. AAAI Vol. 33, pp. 2267-2273 (2015)
3. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL 2004, pp. 271–278 (2004)
4. Diamantini, C., Mircoli, A., Potena, D.: A negation handling technique for sentiment analysis. In: Proceedings of the Collaboration Technologies and Systems (CTS) 2016 (2016)
5. Diamantini, C., Mircoli, A., Potena, D., Storti, E.: Semantic disambiguation in a social information discovery system. In: Proceedings of the Collaboration Technologies and Systems (CTS) 2015, pp. 326-333 (2015)
6. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Engineering Journal, Volume 5, Issue 4, pp. 1093–1113 (2014)
7. Qiu, G., He, X., Zhang, F., Shi, Y., Bu, J., Chen, C.: DASA: dissatisfaction-oriented advertising based on sentiment analysis. Expert Systems with Applications 2010, Vol 37, pp.6182–91 (2010)
8. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proc. of Empirical methods in natural language processing, Volume 10, pp. 79-86 (2002)
9. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proc. of LREC. Genoa, vol. 10, pp. 2200-2204 (2010)
10. Mohammad, S., Turney, P.: Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence, 29 (3), pp. 436-465 (2013)
11. Livingstone, S.R., Thompson, W.F., Wanderley, M.M., Palmer, C.: Common cues to emotion in the dynamic facial expressions of speech and song. The Quarterly Journal of Experimental Psychology, 68:5, pp. 952-970 (2015)
12. Ekman, P.: An argument for basic emotions. Cognition & emotion, 6(3-4), pp. 169-200 (1992)
13. Ekman, P., Sorenson, R.E., Friesen, W.V.: Pan-cultural elements in facial displays of emotions. Science, 164, pp. 86-88 (1969)
14. Tamilselvi, A., ParveenTaj., M.: Sentiment Analysis of Micro blogs using Opinion Mining Classification Algorithm. International Journal of Science and Research (IJSR) 2.10, pp. 196-202 (2013)
15. Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., Perera, A.: Opinion mining and sentiment analysis on a twitter data stream. In: Proc. of advances in ICT for emerging regions, pp. 182-188 (2012)
16. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), Vol. 1631 (2013)
17. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (2009)
18. Mohammad, S. M., Kiritchenko, S.: Using hashtags to capture fine emotion categories from tweets. Computational Intelligence (2014)
19. Busso, Carlos, et al: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th international conference on Multimodal interfaces. ACM (2004)
20. Tian, Y., Kanade, T., Cohn, J.: Recognizing Lower Face Action Units for Facial Expression Analysis. In: Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), pp. 484 – 490 (2000)