AUTOMATIC ANNOTATION OF CORPORA FOR EMOTION RECOGNITION THROUGH FACIAL EXPRESSION ANALYSIS

Claudia Diamantini[1], Alex Mircoli[1], Domenico Potena[1], Emanuele Storti[1]

[1]Department of Information Engineering, Università Politecnica delle Marche (Italy)
{c.diamantini, a.mircoli, d.potena, e.storti}@univpm.it

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Introduction

The massive adoption of social networks has made available an unprecedented amount of user-generated content, which may be analyzed in order to determine people's opinions and emotions on a large variety of topics. Research has made many efforts in defining accurate algorithms for the analysis of emotions conveyed by texts, however their performance often relies on the existence of large annotated datasets, whose current scarcity represents a major issue. The manual creation of such datasets represents a costly and time-consuming activity and hence there is an increasing demand for techniques for the automatic annotation of corpora.

Main Contributions

- Definition of a methodology for the automatic emotional annotation of subtitles on the basis of facial expression analysis
- Development of techniques for the selection of video chunks that are promising for emotional text annotation
- Experimental evaluation of the different phases of the proposed methodology

Methodology

The proposed methodology for the automatic emotional annotation of subtitles through facial expression analysis is based on:

- Ekman's theory of six archetypal facial expressions [1], from which each other emotion can be derived through linear composition;
- empirical evidences of strong correlations between speech and facial expressions [2].

The methodology consists of several phases, as depicted in Figure 1. First, a data source is selected, with particular attention to some issues that could preclude the feasibility and/or the accuracy of emotion detection (*Source Selection*). Afterwards, in a video preprocessing phase, subtitles are analyzed to filter out non-relevant chunks according to their estimated polarity (*Filtering*) and then a proper split of videos is performed on the basis of emotion-bearing chunks (*Video Splitting*). Then, facial expressions of people appearing in the selected frames are analyzed by extracting facial landmarks and averaging by video chunk (*Feature Extraction & Aggregation*). Resulting data are

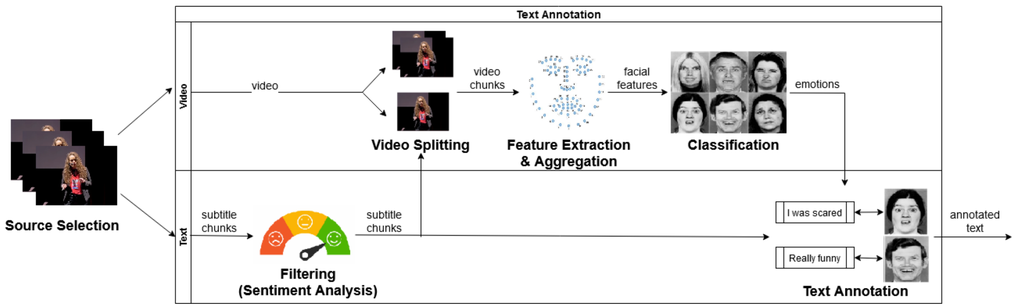

Figure 1. The methodology for the emotional annotation of text

classified through traditional supervised learning techniques (*Classification).* Finally, the resulting emotions are assigned to the corresponding subtitle chunk (*Text Annotation*).

Results

Correlation between opinions and emotions: the proposed methodology is based on the hypothesis that there is a correlation between a non-neutral sentiment polarity and the presence of emotions. For this reason, we empirically evaluated such correlation by considering the test data of the SemEval-2007 Task #14 dataset. The analyzed dataset is composed of 1000 headlines annotated by human operators with respect to both emotions and polarity. As a result, two labels were assigned to each sentence: neutral/non-neutral polarity and neutral/non-neutral emotion. The resulting label distribution is shown in Table I. The correlation between non-neutral sentiment polarity and emotions is equal to 0.92. It means that a sentence with non-neutral polarity usually conveys an emotional content, which supports our preliminary hypothesis.

	Non-neutral polarity	Neutral polarity
Non-neutral emotion	753	136
Neutral emotion	67	21

Table I. Empirical evaluation of the correspondence between sentence polarity and emotions

Facial expression recognition: We considered four emotions (anger, happiness, neutral and sadness) and we manually created a dataset of 200 annotated video chunks with uniform class distribution. For what concerns the final classification layer, we tested Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Decision Tree (DT), Random Forests (RF) and Multi-Layer Perceptron (MLP). We validated each experiment through a 10-fold cross validation. Table II shows the experimental results for the selected classifiers: the SVM classifier outperforms the other models, with a remarkable +8% improvement in accuracy with respect to MLP. The model has an average recall of 0.72, while the average precision is 0.76 and F1=0.74.

Algorithm	Accuracy
SVM	72%
k-NN	58.5%
DT	48%
RF	50%
MLP	64%

Table II. Classifier accuracy evaluated through 10-fold cross validation

Text annotation: we compared the human and automatic text annotations of the dataset composed of 200 video chunks. The results are shown in Table II. The overall accuracy is 64.5%. The major component of the error is represented by misclassification of facial expressions in videos, usually due to the alterations of the facial muscles induced by phonatory movements. We also observed a number of sentences where presumed misclassifications can be actually attributed to wrong human annotations.

	Actual Neutral	Actual Happiness	Actual Anger	Actual Sadness
Predicted Neutral	40	9	13	14
Predicted Happiness	7	38	3	9
Predicted Anger	1	2	26	2
Predicted Sadness	2	1	8	25
Recall	0.80	0.76	0.52	0.50
Precision	0.53	0.67	0.84	0.69

Table III. Confusion matrix for text annotation

Conclusions

- Goal: the definition of a methodology for the automatic creation of annotated corpora through the analysis of facial expressions in subtitled videos.
- The methodology is composed of several video preprocessing phases, with the purpose of filtering out irrelevant frames, and the facial expression classification on the basis of Action Units (AUs) and facial points distances.
- Experiments on a dataset of YouTube videos show that the proposed features for facial expression analysis lead to a 72% accuracy on 4-class emotion recognition.
- Text annotation has a 64.5% accuracy and in some cases the automatic annotation has proven to be able to detect emotions better than human annotators.
- Results could be improved by adopting a multi-modal approach, that includes speech-based emotion features like the tone of the voice.
- Speech recognition techniques could help delimiting frames related to each word, hence implementing a word-level annotation, instead of a chunk-level annotation.

References

[1] P. Ekman, "An argument for basic emotions," Cognition and emotion, vol. 6, pp. 169–200, 1992.
[2] S. Livingstone, W. Thompson, M. Wanderley, and C. Palmer, "Common cues to emotion in the dynamic facial expressions of speech and song," The Quarterly Journal of Experimental Psychology, vol. 68, pp. 952–970, 2015.