Multimodal Sentiment Analysis of Spanish Online Videos

Verónica Pérez Rosas, Rada Mihalcea, Louis-Philippe Morency

Abstract—The number of videos available online and elsewhere is continuously growing, and with this the need for effective methods to process the vast amount of multimodal information shared through this media. This paper addresses the task of multimodal sentiment analysis, and presents a method that integrates linguistic, audio, and visual features for the purpose of identifying sentiment in online videos. We focus our experiments on a new dataset consisting of Spanish videos collected from the social media website YouTube and annotated for sentiment polarity. Through comparative experiments, we show that the joint use of visual, audio, and textual features greatly improves over the use of only one modality at a time. Moreover, we also test the portability of our multimodal method, and run evaluations on a second dataset of English videos.

Index Terms—sentiment analysis, multimodal natural language processing

1 Introduction

Sentiment analysis focuses on the automatic identification of opinions, emotions, evaluations, and judgments, along with their polarity (positive or negative). To date, a large number of applications have already used techniques for automatic sentiment analysis, including branding and product analysis [1], expressive text-to-speech synthesis [2], tracking sentiment timelines in on-line forums and news [3], analysis of political debates [4], question answering [5], conversation summarization [6].

Much of the work to date on sentiment analysis has focused on textual data, and a number of resources have been created including lexicons [7], [8] or large annotated datasets [9], [10]. Given the accelerated growth of other media on the Web and elsewhere, which includes massive collections of videos (e.g., YouTube, Vimeo, VideoLectures), images (e.g., Flickr, Picasa), audio clips (e.g., podcasts), the ability to address the identification of opinions in the presence of diverse modalities is becoming increasingly important.

In this paper, we address the task of multimodal sentiment analysis. We experiment with several linguistic, audio, and visual features, and show that the joint use of these three modalities improves significantly over the use of one modality at a time. As illustrated in Figure 1, modalities other than language can often be used as clues for the expression of sentiment. Their use brings significant advantages over language alone, including: (1) linguistic disambiguation: audio-visual features can help disambiguate the linguistic meaning (e.g., *bomb*); (2) linguistic sparsity problem: audio and visual features bring additional sentiment information (3) grounding: the visual and audio modalities enhance the connection to real world environments [11].

Our main experiments are run on a collection of Spanish videos, with the choice of a language other than English being motivated by the fact that only 27% of Internet users speak English,[1] and the construction of resources and tools for subjectivity and sentiment analysis in languages other than English has been noticed as a growing need [12]. Nonetheless, we also test the portability of our multimodal method, and run evaluations on a second dataset of English videos.

The paper is organized as follows. We first review related work on sentiment and emotion analysis, and then introduce the new dataset of Spanish videos that we use in the main experiments, including a description of the data acquisition, transcription, and annotation. We then present our framework for multimodal sentiment analysis, including a description of the linguistic, audio, and visual features used to build the sentiment classifier. Finally, we present our experiments and evaluations, and conclude with a discussion of the results and a portability experiment on a second English dataset.

2 Related Work

In this section, we review previous work related to multimodal sentiment analysis. We first focus on the problem of text-based sentiment analysis, which has been studied extensively in the field of computational linguistics, and then review the work in audio-visual emotion recognition from the fields of speech processing and computer vision.

2.1 Text-based Sentiment Analysis

The techniques developed so far for sentiment analysis have focused primarily on the processing of text, and

- *Verònica Pérez Rosas and Rada Mihalcea are with the Computer Science Department, University of North Texas, USA, veronica.perezrosas@gmail.com, rada@cs.unt.edu.*
- *Louis-Philippe Morency is with the Institute for Creative Technology, University of Southern California, USA, morency@ict.usc.edu.*

1. www.internetworldstats.com/stats.htm, Oct 11, 2011

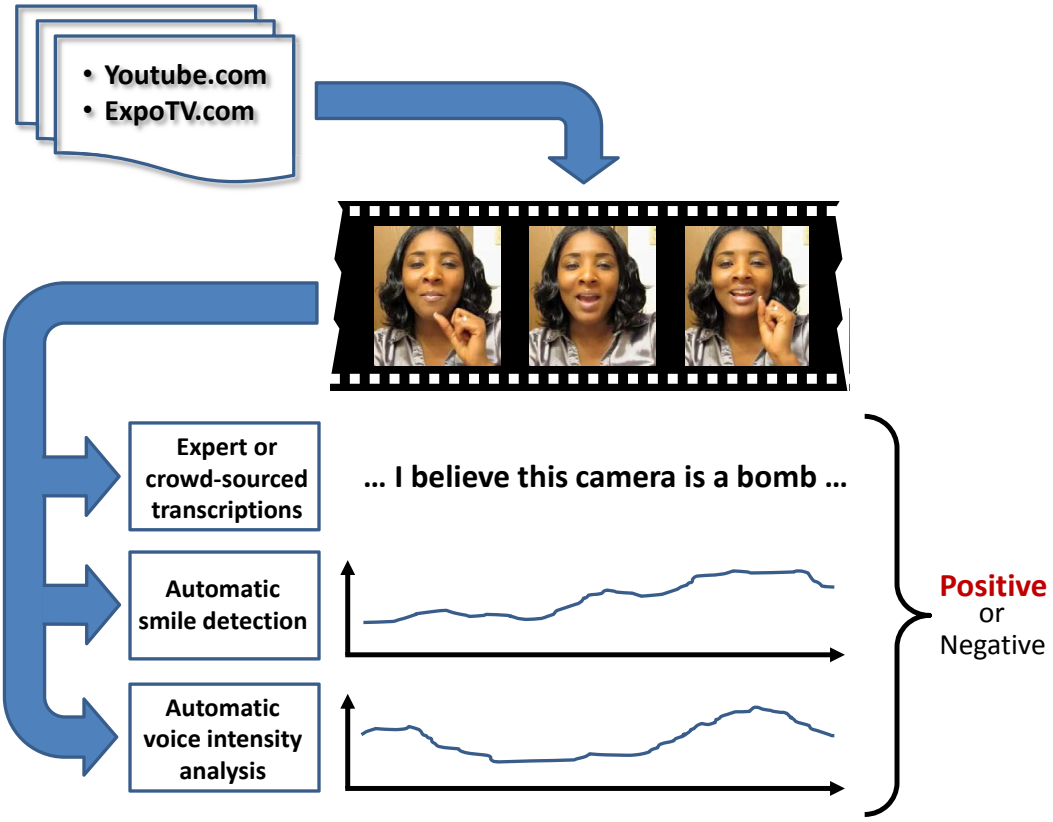

Fig. 1. Overview of our multimodal sentiment analysis approach. The figure shows an example where audio-visual cues help disambiguate the polarity of the spoken utterance. By properly integrating all three sources of information, our approach can successfully recognized the expressed sentiment.

consist of either rule-based classifiers that make use of opinion lexicons, or data-driven methods that assume the availability of a large dataset annotated for polarity.

One of the first lexicons that has been used in polarity analysis is the General Inquirer [13]. Since then, many methods have been developed to automatically identify opinion words [14], [15], as well as n-gram and more linguistically complex phrases [16], [17]. For data-driven methods, one of the most widely used datasets is the MPQA corpus [18], which is a collection of news articles manually annotated for opinions. Other datasets are also available, including two polarity datasets covering the domain of movie reviews [9], [10], and a collection of newspaper headlines annotated for polarity [19]. More recently, multi-domain [20] and multi-lingual [21] resources have also been made available.

Building upon these or other related resources, there is a growing body of work concerned with the automatic identification of subjectivity and sentiment in text, which often addresses online text such as reviews [9], [14], news articles [22], blogs [23], or twitter [24]. Tasks such as cross-domain [20] or cross-language [21], [25] portability have also been addressed. Despite the progress made on the processing of sentiment in text, not much has been done in terms of extending the applicability of sentiment analysis to other modalities, such as speech, gesture, or facial expressions. We are only aware of two exceptions. First, in the research reported in [26], speech and text are analyzed jointly for the purpose of subjectivity identification. This previous work, however, did not address other modalities such as visual cues, and did not address the problem of sentiment analysis. More recently, in a pre-study on 47 English videos [27], it has been shown that visual and audio features can complement textual features for sentiment analysis. In our work, we use a new dataset focusing on Spanish, and draw summary features at video level. Moreover, we show that multimodal sentiment analysis can be effectively used for sentiment analysis on different languages.

2.2 Audio-Visual Emotion Analysis

Over the past few years, we have seen a new line of research addressing the multimodal fusion of language, acoustic features, and visual gestures, such as the VIRUS project that uses all three modalities to perform video retrieval [28].

Along these lines, closely related to our work is the research done on audio and/or visual emotion analysis. For recent surveys of dimensional and categorical affect recognition see Zeng *et al.* [29], and Gunes and Pantic [30]. For instance, a novel algorithm is defined in [31], based

on a combination of audio-visual features for emotion recognition. Nicolaou *et al.* [32] propose the use of Output-Associative Relevance Vector Machine (OA-RVM) for dimensional and continuous prediction of emotions based on automatically tracked facial feature points.

In addition to work that considered individual audio or visual modalities [33], [34], [35], there is also a growing body of work concerned with audio-visual emotion analysis [36], [37], [38]. The features used by these novel algorithms are usually low level features, such as tracking points for collecting visual data, or audio features like pitch level. More recently, a challenge has been organized focusing on the recognition of emotions using audio and visual cues [39], [40], which included sub-challenges on audio-only, video-only, and audio-video, and drew the participation of many teams from around the world. Also related to our work is the multimodal integration of opinion mining and facial expressions, which can be successfully used for the development of intelligent affective interfaces [41].

It is also important to note that multimodal emotion recognition is different from multimodal sentiment analysis. While opinion polarity may often be correlated to emotional valence (as used for instance in the datasets for audio-video emotion analysis [39]), these concepts are very different. For instance, someone can be smiley while at the same time expressing a negative opinion, which makes multimodal sentiment analysis a different and yet challenging research direction.

3 A Spanish Multimodal Opinion Dataset

We collect a new dataset consisting of 105 videos in Spanish from the social media web site YouTube. An important characteristic of our dataset is its generalized nature; the dataset is created in such a way that it is not based on one particular topic. The videos were found using the following keywords: *mi opinion* (my opinion), *mis products favoritos* (my favorite products), *me gusta* (I like), *no me gusta* (I dislike), *products para bebe* (baby products), *mis perfumes favoritos* (my favorite perfumes), *peliculas recomendadas* (recommended movies), *opinion politica* (politic opinion), *video juegos* (video games) and *abuso animal* (animal abuse). To select the videos, we used the following guidelines: people should be in front of the camera; their face should be visible; there should not be any background music or animation. Figure 2 shows example snapshots of our dataset.

The final video set includes 21 male and 84 female speakers randomly selected from YouTube, with their age approximately ranging from 15 to 60 years. Although from different Spanish speaking countries (e.g., Spain, Mexico, or countries from South America), all speakers expressed themselves in Spanish. The videos are converted into the *mp4* format with a standard size of 352x288. The length of the videos varies from 2-8 minutes.

Fig. 2. Example snapshots from out Spanish Multimodal Opinion Dataset.

All videos are pre-processed to address the following issues: introductory titles and multiple topics. Many videos on YouTube contain an introductory sequence where a title is shown, sometimes accompanied with a visual animation. As a simple way to address this issue, we manually segment the video until the beginning of the first opinion utterance. In the future we are planning to optimize this by automatically performing optical character (OCR) and face recognition on the videos [42].

The second issue is related to multiple topics. Video reviews can address more than one topic (or aspect). For example, a person can start by talking about the food served in the restaurant and then switch to a new topic about eating habits. To simply address this issue, all video sequences are normalized to be about 30 seconds in length, while making sure that no utterances are cut half way. We keep as future work to automatically segment topics based on transcriptions [43] or directly based on the audio-visual signals.

3.1 Transcriptions

All video clips are manually transcribed to extract spoken words as well as the start and end time of each spoken utterance. The Transcriber software was used to perform this task. The transcription is done using only the audio track without visual information. Each video contains from 4 to 12 utterances, with most videos having 6-8 utterances in the extracted 30 seconds. The utterance segmentation was based on long pauses that could easily be detected using tools such as Praat and OpenEAR [44]. The final set of transcriptions contains approximately 550 utterances / 10,000 words.

Multimodal sentiment analysis using manual transcription is a precedent step to fully automatic sentiment classification. Manual transcription and segmentation are very reliable but also time consuming. Alternatives for performing the transcription step automatically include the use of automatic speech recognition, with technologies such as Google voice or Adobe translator, or the use of crowd-sourcing techniques such as Amazon Mechanical Turk. In the Results and Discussion section, we present an English dataset which was efficiently transcribed using this crowd-sourcing approach.

3.2 Sentiment Annotations

Since our goal is to automatically find the sentiment expressed in the video clip, we decided to perform our annotation task at video sequence level. This is an important step while creating the dataset and we were particularly careful while describing the task. We ask the annotators to associate a sentiment label that best summarizes the opinion expressed in the YouTube video and not the sentiment felt while watching the video.

For each video, we want to assign one of these three labels: negative, neutral, or positive. All 105 video clips were annotated by two annotators who were shown videos in two random sequencing orders. The average interannotator agreement is 92%, with a κ of 0.84, which indicates strong agreement. To determine the final gold standard label, all the annotation disagreements were resolved though discussion. The final dataset consists of 105 video clips, out of which 47 are labeled as positive, 54 as negative, and 4 as neutral. The baseline on this dataset is 51%, which corresponds to the accuracy obtained if all the videos are assigned with the most frequent polarity label in the dataset.

4 Multimodal Sentiment Analysis

The greatest advantage of analyzing video opinions as compared to text-only opinions is that additional cues can be used. In textual opinions, the only available source of information consists of the words in the opinion and the dependencies among them, which may sometime prove insufficient to convey the exact sentiment of the consumer. Instead, video opinions provide multimodal data in the form of vocal as well as visual responses. The vocal modulations in the recorded response help us determine the tone of the speaker whereas visual data can provide information regarding the emotional state of the speaker. Thus our hypothesis is that a combination of text and video data can help create a better analysis model. We specifically focus on the three main types of features covering the three modalities.

4.1 Linguistic Features

We use a bag-of-words representation of the video transcriptions to derive unigram counts, which are then used as input features. First, we build a vocabulary consisting of all the words, including stopwords, occurring in the transcriptions of the training set. We then remove those words that have a frequency below 10 (value determined empirically on a small development set). The remaining words represent the unigram features, which are then associated with a value corresponding to the frequency of the unigram inside each transcription. These simple weighted unigram features have been successfully used in the past to build sentiment classifiers on text, and in conjunction with Support Vector Machines have been shown to lead to state-of-the-art performance [9], [10].

4.2 Audio Features

The audio features are automatically extracted from the audio track of each video clip. The audio features are extracted at the same frame rate as the video features (30Hz) with a sliding window of 50ms. We used the open source software OpenEAR [44] to automatically compute the pitch and voice intensity. Speaker normalization is performed using z-standardization. The voice intensity was simply thresholded to identify samples with and without speech. The same threshold was used for all the experiments and all the speakers.

For each video in our dataset, we define four summary features:

- **Pause duration:** Given the audio frames extracted from the entire video, how many audio samples are identified as silence. This audio feature is then normalized by the number of audio samples in the video. This feature can be interpreted as the percentage of the time where the speaker was silent.
- **Pitch:** Compute the standard deviation of the pitch level for the video. This measure represents the variation of voice intonation during the entire video.
- **Intensity:** Measure the sound power of the spoken utterances in the video. We compute the average voice intensity over the whole video.
- **Loudness:** Determine the perceived strength of the voice factored by the ear's sensitivity. We compute the average loudness measure over the entire video.

4.3 Visual Features

The visual features are automatically extracted from the video sequences. Since only one person is present in each video clip and they are most of the time facing the camera, current technology for facial tracking can efficiently be applied to our dataset. We use a commercial software called OKAO Vision that detects at each frame the face, it extracts the facial features, and extrapolates some basic facial expressions as well as eye gaze direction. The main facial expression being recognized is smile. This is a well-established technology that can be found in many digital cameras. For each frame, the vision software returns a smile intensity (0-100) and the gaze direction, using both horizontal and vertical angles expressed in degrees. The sampling rate is the same as the video framerate: 30Hz.

An important aspect when generating visual features is the quality of the video, and correspondingly the quality of the visual processing that can be automatically performed on the video. OKAO provides a confidence level for each processed frame in the range 0-1000. We discounted all the frames with a confidence level below 700, and we also removed all the videos for which more than 30% of the frames had a confidence below 700.

For each video in our dataset, we define two series of summary features:

- **Smile duration:** Given all the frames in a video, how many frames are identified as "smile." In our

Modality	Accuracy
Text only	64.94%
Visual only	61.04%
Audio only	46.75%
Text-visual	73.68%
Text-audio	68.42%
Audio-visual	66.23%
Text-audio-visual	**75.00%**

TABLE 1
Automatic sentiment classification performance for seven different models on our Spanish multimodal opinion dataset. One modality at a time: text-only, visual-only, audio-only; two modalities at a time: text-visual, text-audio, visual-audio; all three modalities: text-audio-visual.

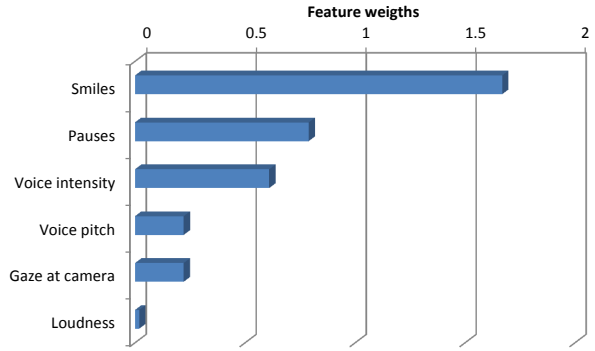

Fig. 3. Visual and audio feature weights. This graph shows the relative importance of the SVM weights associated to each audio-visual features.

experiments, we use three different variants of this feature with different thresholds: 50 and 75.

- **Look-away duration:** Given all the frames in a video, in how many frames is the speaker looking at the camera. The horizontal and vertical angular thresholds were experimentally set to 10 degrees.

The visual features are normalized by the total number of frames during the video. Thus, if the person is smiling half the time, then the smile feature will be equal to 0.5 (or 50%).

5 Experiments

Our main experiments are run on the new Spanish multimodal opinion dataset introduced earlier. From the dataset, we remove those videos that have low visual processing performance (i.e., the number of frames correctly processed by OKAO below 70%), and further remove the videos labeled as neutral (i.e., keeping only positive and negative videos). This leaves us with an experimental dataset of 76 videos, consisting of 39 positive and 37 negative videos, for which we extract linguistic, audio, and visual features as described above.

The multimodal fusion is performed using the early fusion technique, where all the linguistic, audio, and visual features are concatenated into a common feature vector, thus resulting in one vector for each video in the dataset. For the classification, we use Support Vector Machines (SVM) with a linear kernel, which are binary classifiers that seek to find the hyperplane that best separates a set of positive examples from a set of negative examples, with maximum margin [45]. We use the Weka machine learning toolkit. For each experiment, a ten-fold cross validation is run on the entire dataset.

6 Results and Discussion

Table 1 shows the results obtained with one, two, and three modalities at a time. The experiments performed on the newly introduced dataset of Spanish videos show that the integration of visual, audio, and textual features can improve significantly over the individual use of one modality at a time. Among the individual classifiers, the text classifier appears to be the most accurate, followed by the classifier that relies on visual clues, and by the audio classifier.

6.1 Feature Analysis

To determine the role played by each of the visual and audio features, we compare the feature weights assigned by the SVM learning algorithm, as shown in Figure 3. Perhaps not surprisingly, the smile is the most predictive feature, followed by the number of pauses and voice intensity. Voice pitch, gaze at camera, and loudness are also contributing to the classification, but to a lesser extent.

To determine how these features affect the polarity classification, Figure 4 shows the average values calculated for the three most predictive features: smiles, pauses, and voice intensity. As seen in this figure, an increased number of smiles and an increased number of pauses are characteristic for positive videos, whereas higher voice intensity is more typical for negative videos. It thus appears that the speakers of a negative review would have a higher voice intensity and speak at a higher rate (i.e., pause less), unlike the speakers of a positive review who tend to be at a slower pace when they speak.

6.2 Multimodal sentiment analysis on English videos

As a final experiment, to determine the portability of the multimodal sentiment analysis method to a different dataset, we compile a second dataset consisting of English video reviews. We collect cellular phone reviews from ExpoTv,[2] which is a public website that provides consumer generated videos. Through this platform users provide unbiased video opinions of products organized in various categories. We started by collecting 37 reviews, which were then filtered using the same criteria as used to build the Spanish dataset. One additional challenge that we faced in this dataset is occlusion, with people

2. http://www.expotv.com

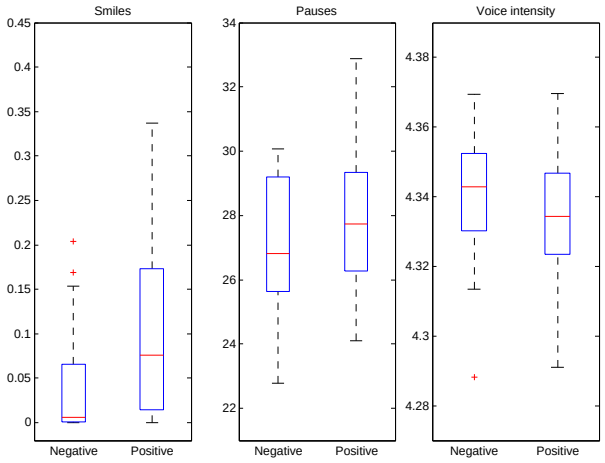

Fig. 4. Average values of several multimodal features when clustered per sentiment label.

Modality	**Accuracy**
Text only	54.05%
Visual only	54.05%
Audio only	48.64%
Text-audio-visual	**64.86%**

TABLE 2
Multimodal sentiment analysis on an English dataset.

often showing the product they review to the camera, thus covering their face. Since our visual processing approach is applied independently on each frame, images with occluded faces were simply ignored during the summary feature calculations. As before, from each video we manually extract a 30 seconds segment in which people express their opinion.

To obtain the transcriptions, for this dataset we experiment with a different technique, and use crowd-sourcing via Amazon Mechanical Turk. To ensure quality, the transcriptions collected from the Amazon service were verified by one of the authors of this paper. For the sentiment annotations, since ExpoTv users provide a star rating to the product they are reviewing (one to five stars), we use this information to assign a sentiment label to each video: videos with four or five stars are labeled as positive, whereas videos with one or two stars are labeled as negative. Using this labeling approach, we ended up with 20 positive reviews and 17 negative reviews.

Table 2 shows the results obtained on this dataset. As before, the joint use of all three modalities brings significant improvements over models that use only one modality at a time. Interestingly, similar to the experiments performed on the Spanish dataset, the audio model is the weakest model, which suggests audio feature engineering as a possible avenue for future work.

7 Conclusions

In this paper, we addressed the task of multimodal sentiment analysis, and explored the joint use of multiple modalities for the purpose of classifying the polarity of opinions in online videos. Through experiments performed on a newly introduced dataset, consisting of Spanish videos where people express their opinion about different topics, we showed that the integration of visual, audio, and textual features can improve significantly over the individual use of one modality at a time. Moreover, we also tested the portability of our multimodal method, and showed that significant improvements are also obtained on a second dataset of English videos. While additional research is needed to explore datasets covering other domains and languages, we believe our initial experiments show the promise of this research direction.

The datasets introduced in this paper are available upon request.

Acknowledgments

This material is based in part upon work supported by National Science Foundation awards #0917170 and #0917321. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, Washington, 2004.

[2] C. Alm, D. Roth, and R. Sproat, "Emotions from text: Machine learning for text-based emotion prediction," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Vancouver, Canada, 2005, pp. 347–354.

[3] L. Lloyd, D. Kechagias, and S. Skiena, "Lydia: A system for large-scale news analysis," in *String Processing and Information Retrieval (SPIRE 2005)*, 2005.

[4] P. Carvalho, L. Sarmento, J. Teixeira, and M. Silva, "Liars and saviors in a sentiment annotated corpus of comments to political debates," in *Proceedings of the Association for Computational Linguistics (ACL 2011)*, Portland, OR, 2011.

[5] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in *Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, Sapporo, Japan, 2003, pp. 129–136.

[6] G. Carenini, R. Ng, and X. Zhou, "Summarizing emails with conversational cohesion and subjectivity," in *Proceedings of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2008)*, Columbus, Ohio, 2008.

[7] J. Wiebe and E. Riloff, "Creating subjective and objective sentence classifiers from unannotated texts," in *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005) (invited paper)*, Mexico City, Mexico, 2005.

[8] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, Genova, IT, 2006.

[9] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, Barcelona, Spain, July 2004.

[10] A. Maas, R. Daly, P. Pham, D. Huang, A. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the Association for Computational Linguistics (ACL 2011)*, Portland, OR, 2011.

[11] D. Chen and R. Mooney, "Panning for gold: Finding relevant semantic content for grounded language learning," in *Proceedings of Symposium on Machine Learning in Speech and Language Processing*, 2011.

[12] C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual sentiment and subjectivity," in *Multilingual Natural Language Processing*. Prentice Hall, 2011.
[13] P. Stone, *General Inquirer: Computer Approach to Content Analysis*. MIT Press, 1968.
[14] P. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, 2002, pp. 417–424.
[15] M. Taboada, J. Brooke, M. Tofiloski, K. Voli, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 3, 2011.
[16] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, 2003, pp. 105–112.
[17] H. Takamura, T. Inui, and M. Okumura, "Latent variable models for semantic orientations of phrases," in *Proceedings of the European Chapter of the Association for Computational Linguistics*, 2006.
[18] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2-3, pp. 165–210, 2005.
[19] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007)*, Prague, Czech Republic, 2007.
[20] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Association for Computational Linguistics*, 2007.
[21] C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual subjectivity: Are more languages better?" in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China, August 2010, pp. 28–36. [Online]. Available: http://www.aclweb.org/anthology/C10-1004
[22] K. Balog, G. Mishne, and M. de Rijke, "Why are they excited? identifying and explaining spikes in blog mood levels," in *Proceedings of the 11th Meeting of the European Chapter of the As sociation for Computational Linguistics (EACL-2006)*, 2006.
[23] N. Godbole, M. Srinivasaiah, and S. Sekine, "Large-scale sentiment analysis for news and blogs," in *International Conference on Weblogs and Social Media*, Denver, CO, 2007.
[24] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification," in *Proceedings of the Association for Computational Linguistics (ACL 2011)*, Portland, OR, 2011.
[25] X. Wan, "Co-training for cross-lingual sentiment classification," in *Proceedings of the Joint Conference of the Association of Computational Linguistics and the International Joint Conference on Natural Language Processing*, Singapore, August 2009.
[26] S. Raaijmakers, K. Truong, and T. Wilson, "Multimodal subjectivity analysis of multiparty conversation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, 2008, pp. 466–474.
[27] L. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the International Conference on Multimodal Computing*, Alicante, Spain, 2011.
[28] P. Martins, T. Langlois, and T. Chambel, "Movieclouds: Content-based overviews and exploratory browsing of movies," in *Proceedings of the Academic MindTrek*, Tampere, Finland, 2011.
[29] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.
[30] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int'l Journal of Synthetic Emotion*, vol. 1, no. 1, pp. 68–99, 2010.
[31] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, October 2010.
[32] M. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," in *IEEE FG'11*, 2011.
[33] M. E. Hoque, R. el Kaliouby, and R. Picard, "When human coders (and machines) disagree on the meaning of facial affect in spontaneous videos," in *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, Amsterdam, Netherlands, 2009.
[34] Y. Shin, Y. Kim, and E. Kim, "Automatic textile image annotation by predicting emotional concepts from visual features," *Image and Vision Computing*, vol. 28, no. 3, 2010.
[35] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, 2003.
[36] N. Sebe, I. Cohen, T. Gevers, and T. Huang, "Emotion recognition based on joint visual and audio cues," in *ICPR*, 2006.
[37] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2331–2347, November 2007.
[38] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Transactions on Multimedia*, vol. 10, no. 5, 2008.
[39] B. Schuller, M. Valstar, R. Cowie, and M. Pantic, Eds., *Audio/Visual Emotion Challenge and Workshop (AVEC 2011)*, 2011.
[40] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic, Eds., *Audio/Visual Emotion Challenge and Workshop (AVEC 2011)*, 2011.
[41] E. Cambria, I. Hupont, A. Hussain, E. Cerezo, and S. Baldassarri, *Sentic Avatar: Multimodal Affective Conversational Agent with Common Sense*. Springer Book Series, 2011.
[42] T. Plotz and G. A. Fink, "Markov models for offline handwriting recognition: a survey," *International Journal on Document Analysis and Recognition*, vol. 12, no. 4, 2009.
[43] J. Arguello and C. Rose, "Topic segmentation of dialogue," in *HLT-NAACL Workshop on Analyzing Conversations in Text and Speech*, 2009.
[44] F. E. M. W. B. Schuller, "Openear introducing the munich open-source emotion and affect recognition toolkit," in *ACII*, 2009.
[45] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

Verónica Pérez Rosas is a doctoral student in the Department of Computer Science and Engineering at the University of North Texas. She received her Master degree in Computer Science from the Instituto Tecnologico de Cd. Madero in 2008. Her research interests include Natural Language Processing, Machine Learning, and Intelligent Optimization. Her current research is centered around methods for subjectivity and sentiment analysis, where she specifically focuses on the role played by different data modalities (e.g., text, audio and video) for the task of sentiment analysis. She has authored papers in both optimization and Natural Language Processing conferences, and serves or has served as a reviewer for several conferences.

Rada Mihalcea is an Associate Professor in the Department of Computer Science and Engineering at the University of North Texas. Her research interests are in computational linguistics, with a focus on lexical semantics, graph-based algorithms for natural language processing, and multilingual natural language processing. She serves or has served on the editorial boards of the Journals of Computational Linguistics, Language Resources and Evaluations, Natural Language Engineering, Research in Language in Computation, IEEE Transations on Affective Computing, and Transactions of the Association for Computational Linguistics. She was a program co-chair for the Conference of the Association for Computational Linguistics (2011), and the Conference on Empirical Methods in Natural Language Processing (2009). She is the recipient of a National Science Foundation CAREER award (2008) and a Presidential Early Career Award for Scientists and Engineers (2009).

Louis-Philippe Morency is a Research Assistant Professor in the Department of Computer Science at the University of Southern California (USC) and Research Scientist at the USC Institute for Creative Technologies where he leads the Multimodal Communication and Machine Learning Laboratory (MultiComp Lab). He received his Ph.D. and Master degrees from MIT Computer Science and Artificial Intelligence Laboratory. His research interests are in computational study of nonverbal social communication, a multi-disciplinary research topic that overlays the fields of multimodal interaction, computer vision, machine learning, social psychology and artificial intelligence. Dr. Morency was selected in 2008 by IEEE Intelligent Systems as one of the Ten to Watch for the future of AI research. He received 6 best paper awards in multiple ACM- and IEEE-sponsored conferences for his work on context-based gesture recognition, multimodal probabilistic fusion and computational modeling of human communication dynamics. His work was reported in The Economist, New Scientist and Fast Company magazines.