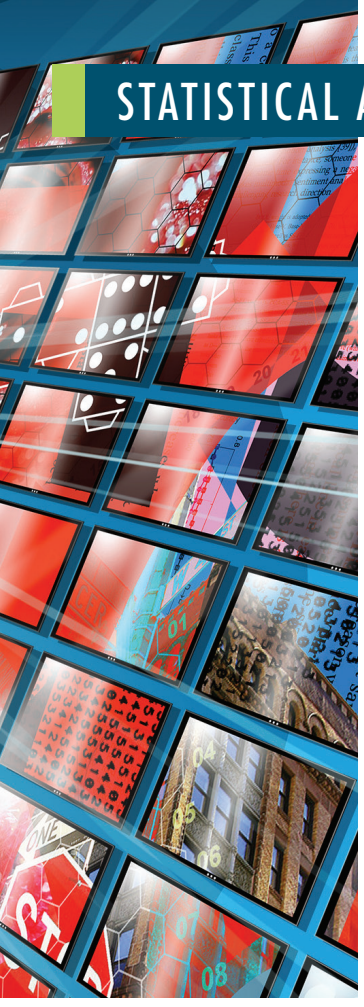

Multimodal Sentiment Analysis of Spanish Online Videos

Verónica Pérez Rosas, *University of North Texas*

Rada Mihalcea, *University of Michigan*

Louis-Philippe Morency, *University of Southern California*

Using multimodal sentiment analysis, the presented method integrates linguistic, audio, and visual features to identify sentiment in online videos. In particular, experiments focus on a new dataset consisting of Spanish videos collected from YouTube that are annotated for sentiment polarity.

Sentiment analysis focuses on the automatic identification of opinions, emotions, evaluations, and judgments, along with their polarity (positive or negative). Techniques for automatic sentiment analysis are already in use in many applications, including those related to branding and product analysis,[1] expressive text-to-speech synthesis,[2] tracking sentiment timelines in online forums and news,[3] analyzing political debates,[4] answering questions,[5] and summarizing conversations.[6]

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

Here, we address the task of multimodal sentiment analysis (for other research approaches, see the sidebar "Related Work in Multimodal Sentiment Analysis"). We experiment with several linguistic, audio, and visual features, and show that the joint use of these three modalities significantly improves the classification accuracy, as compared to using one modality at a time. As Figure 1 shows, modalities other than language can often be used as clues for the expression of sentiment. Their use brings significant advantages over language alone, including

- *linguistic disambiguation*—audio-visual features can help disambiguate linguistic meaning (for example, the word *bomb*);
- *linguistic sparsity problem*—audio and visual features bring additional sentiment information; and
- *grounding*—the visual and audio modalities enhance the connection to real-world environments.[11]

Our main experiments were run on a collection of Spanish videos; we chose a language other than English because only 27 percent of Internet users speak English (www.internetworldstats.com/stats.htm, 11 October 2011), and constructing resources and tools for subjectivity and sentiment analysis

1541-1672/13/$31.00 © 2013 IEEE
Published by the IEEE Computer Society

in languages other than English is a growing need.[12] We also tested the portability of our multimodal method and ran evaluations on a second dataset of English videos.

A Spanish Multimodal Opinion Dataset

We collected a new dataset consisting of 105 videos in Spanish from the social media website YouTube. An important characteristic of our dataset is its generalized nature; the dataset is created in such a way that it's not based on one particular topic. We found the videos 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* (political 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 shouldn't be any background music or animation. Figure 2 shows example snapshots from our dataset.

The final video set includes 21 male and 84 female speakers randomly selected from YouTube, with their age ranging from approximately 15 to 60. Although they're from different Spanish-speaking countries (such as Spain, Mexico, and various South American countries), all of the speakers expressed themselves in Spanish. The videos were converted into the .mp4 format with a standard size of 352 × 288. The length of the videos varies from 2–8 minutes.

We preprocessed all of the videos to address two issues: introductory titles and multiple topics. Many videos on YouTube contain an introductory sequence that shows a title, sometimes accompanied by a visual animation. As a simple way to address this issue, we manually segmented the video until the beginning of the first opinion utterance. In the future, we plan to optimize this by automatically performing optical character recognition (OCR) and face recognition on the videos.[13]

Videos—particularly video reviews—might address more than one topic (or

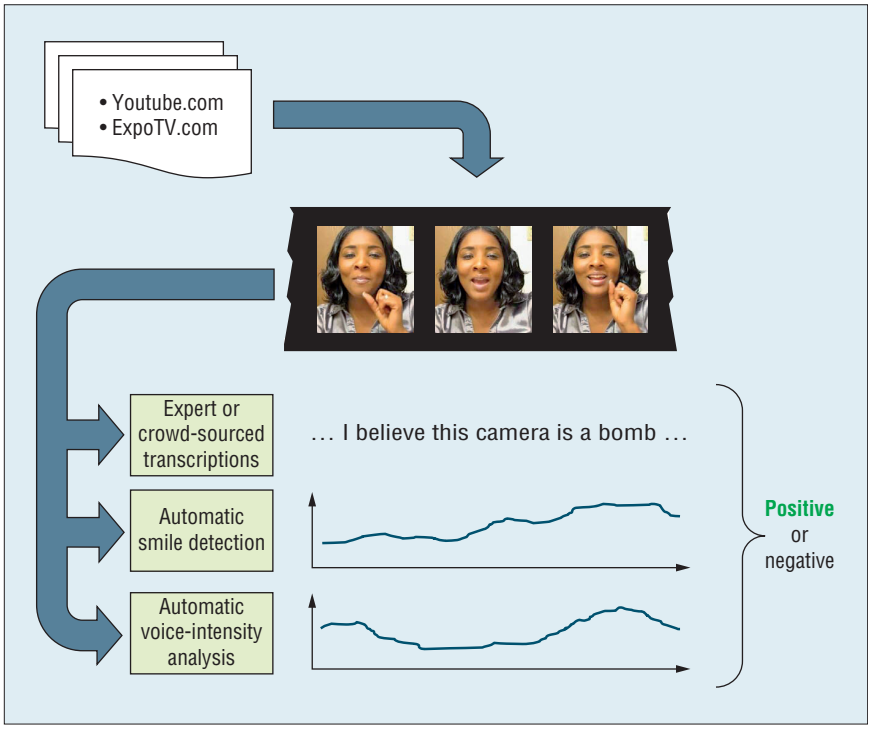

Figure 1. Overview of the multimodal sentiment analysis approach. In this example, audio-visual cues help disambiguate the polarity of the spoken utterance. By properly integrating all three sources of information, this approach can successfully recognize the expressed sentiment.

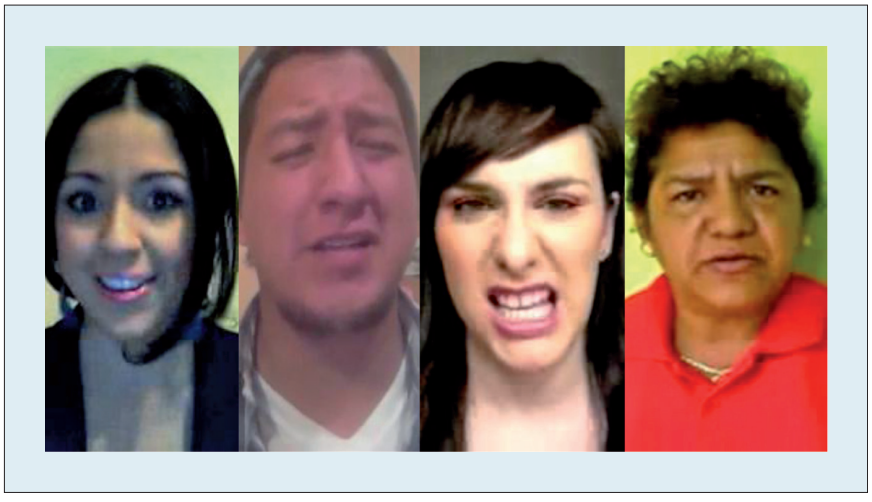

Figure 2. Example snapshots from our Spanish Multimodal Opinion Dataset. In this figure, we can observe differences in facial expressions elicited by people while expressing positive and negative opinions.

Related Work in Multimodal Sentiment Analysis

Our overview of previous work related to multimodal sentiment analysis falls into two categories: text-based sentiment analysis, which has been studied extensively in the field of computational linguistics; and audio-visual emotion recognition from the fields of speech processing and computer vision.

Text-Based Sentiment Analysis

The techniques developed so far for sentiment analysis have focused primarily on text processing and consist of either rule-based classifiers that use opinion lexicons, or data-driven methods that assume the availability of a large dataset annotated for polarity.

The General Inquirer was one of the first lexicons used in polarity analysis;[1] since its introduction, many methods have been developed to automatically identify opinion words,[2,3] *n*-grams, and more linguistically complex phrases.[4,5] For data-driven methods, one of the most widely used datasets is the MPQA corpus,[6] 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,[7,8] and a collection of newspaper headlines annotated for polarity.[9] More recently, multidomain[10] and multilingual[11] resources have also become available.

Building upon these and other related resources, there's a growing body of work concerned with the automatic identification of subjectivity and sentiment in text, which often addresses online text such as reviews,[2,7] news articles,[12] blogs,[13] or Twitter.[14] Tasks such as cross-domain[15] or cross-language[11,15] 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're aware of only two exceptions. First, in research reported elsewhere,[16] speech and text are analyzed jointly for the purpose of subjectivity identification. This previous work, however, didn't address other modalities such as visual cues, and didn't address the problem of sentiment analysis.

More recently, in a pre-study on 47 English videos,[17] 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 the video level. Moreover, we show that multimodal sentiment analysis can be effectively used for sentiment analysis on different languages.

Audio-Visual Emotion Analysis

Over the past few years, we've seen a new line of research addressing the multimodal fusion of language, acoustic features, and visual gestures, such as the Video Information Retrieval Using Subtitles (Virus) project that uses all three modalities to perform video retrieval.[18]

Along these lines—and closely related to our own work—is the research on audio and/or visual emotion analysis. Some recent surveys discuss dimensional and categorical affect recognition.[19,20] For instance, Martin Wollmer and his colleagues define a novel algorithm based on a combination of audio-visual features for emotion recognition.[21] Mihalis Nicolaou and his colleagues 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.[22]

In addition to work that considered individual audio or visual modalities,[23–25] there's also a growing body of work concerned with audio-visual emotion analysis.[26–28] 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 was organized focusing on the recognition of emotions using audio and visual cues,[29] which included subchallenges 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.[30]

It's also important to note that multimodal emotion recognition is different from multimodal sentiment analysis. Although opinion polarity is often correlated to emotional valence (as used, for instance, in the datasets for audio-video emotion analysis[29]), these concepts are quite different. For instance, someone can be smiling while at the same time expressing a negative opinion, which makes multimodal sentiment analysis a complex and challenging research direction.

aspect). For example, a person can start by talking about the food served in a restaurant and then switch to a discussion of eating habits. To address this issue, all video sequences were normalized to be about 30 seconds in length, while ensuring that no utterances were cut off. In future work, we hope to automatically segment topics based on transcriptions[14] or on audio-visual signals directly.

Transcriptions

All video clips were manually transcribed to extract spoken words as well as the start and end time of each spoken utterance. Transcriber software was used to perform this task. The transcription was performed using only the audio track, without visual information. Each video contains 4–12 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.[15] The final set of transcriptions contains approximately 550 utterances and 10,000 words.

Multimodal sentiment analysis using manual transcription is a precedent step to fully automatic sentiment classification. Manual transcription

References

1. P. Stone, *General Inquirer: Computer Approach to Content Analysis*, MIT Press, 1968.
2. P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proc. 40th Ann. Meeting of the Assoc. Computational Linguistics*, ACL, 2002, pp. 417–424.
3. M. Taboada et al., "Lexicon-Based Methods for Sentiment Analysis," *J. Computational Linguistics*, vol. 37, no. 3, 2011, pp. 267–307.
4. E. Riloff and J. Wiebe, "Learning Extraction Patterns for Subjective Expressions," *Proc. Conf. Empirical Methods in Natural Language Processing*, 2003, pp. 105–112.
5. H. Takamura, T. Inui, and M. Okumura, "Latent Variable Models for Semantic Orientations of Phrases," *Proc. European Chapter of the Association for Computational Linguistics*, ACL, 2006, pp. 201–208.
6. J. Wiebe, T. Wilson, and C. Cardie, "Annotating Expressions of Opinions and Emotions in Language," *Language Resources and Evaluation*, vol. 39, nos. 2–3, 2005, pp. 165–210.
7. B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *Proc. 42nd Meeting of the Assoc. Computational Linguistics*, ACL, 2004, pp. 271–278.
8. A. Maas et al., "Learning Word Vectors for Sentiment Analysis," *Proc. Assoc. Computational Linguistics*, ACL, 2011, pp. 142–150.
9. C. Strapparava and R. Mihalcea, "Semeval-2007 Task 14: Affective Text," *Proc. 4th Int'l Workshop on the Semantic Evaluations*, ACL, 2007, pp. 70–74.
10. J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boom-Boxes, and Blenders: Domain Adaptation for Sentiment Classification," *Proc. Assoc. Computational Linguistics*, ACL, 2007, pp. 187–205.
11. C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual Subjectivity: Are More Languages Better?" *Proc. 23rd Int'l Conf. Computational Linguistics*, ACL, 2010, pp. 28–36.
12. K. Balog, G. Mishne, and M. de Rijke, "Why Are They Excited? Identifying and Explaining Spikes in Blog Mood Levels," *Proc. 11th Meeting of the European Chapter of the Assoc. Computational Linguistics*, ACL, 2006, pp. 207–210.
13. N. Godbole, M.Srinivasaiah, and S. Sekine, "Large-Scale Sentiment Analysis for News and Blogs," *Proc. Int'l Conf. Weblogs and Social Media*, 2007; http://icwsm.org/papers/3--Godbole-Srinivasaiah-Skiena.pdf.
14. L. Jiang et al., "Target-Dependent Twitter Sentiment Classification," *Proc. Assoc. Computational Linguistics*, ACL, 2011, pp. 151–160.
15. X. Wan, "Co-Training for Cross-Lingual Sentiment Classification," *Proc. Joint Conf. Assoc. Computational Linguistics and the Int'l Joint Conf. Natural Language Processing*, ACL, 2009, pp. 235–243.
16. S. Raaijmakers, K. Truong, and T. Wilson, "Multimodal Subjectivity Analysis of Multiparty Conversation," *Proc. Conf. Empirical Methods in Natural Language Processing*, ACL, 2008, pp. 466–474.
17. L. Morency, R. Mihalcea, and P. Doshi, "Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web," *Proc. Int'l Conf. Multimodal Computing*, ACM, 2011, pp. 169–176.
18. P. Martins, T. Langlois, and T. Chambel, "Movieclouds: Content-Based Overviews and Exploratory Browsing of Movies," *Proc. Academic MindTrek*, ACM, 2011, pp. 133–140.
19. Z. Zeng et al., "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, 2009, pp. 39–58.
20. H. Gunes and M. Pantic, "Automatic, Dimensional and Continuous Emotion Recognition," *Int'l J. Synthetic Emotion*, vol. 1, no. 1, 2010, pp. 68–99.
21. M. Wollmer et al., "Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening," *IEEE J. Selected Topics in Signal Processing*, vol. 4, no. 5, 2010, pp. 867–881.
22. M. Nicolaou, H. Gunes, and M. Pantic, "Output-Associative RVM Regression for Dimensional and Continuous Emotion Prediction," *Proc. IEEE Automatic Face and Gesture Recognition Workshops*, IEEE, 2011, pp. 16–23.
23. M.E. Hoque, R. el Kaliouby, and R. Picard, "When Human Coders (and Machines) Disagree on the Meaning of Facial Affect in Spontaneous Videos," *Proc. 9th Int'l Conf. Intelligent Virtual Agents*, LNCS 5773, Springer, 2009, pp. 337–343.
24. 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, pp. 526–537.
25. K. Scherer, "Vocal Communication of Emotion: A Review of Research Paradigms," *Speech Comm.*, vol. 40, nos. 1–2, 2003, pp. 227–256.
26. N. Sebe et al., "Emotion Recognition Based on Joint Visual and Audio Cues," *IEEE Int'l Conf. Pattern Recognition*, IEEE, 2006, pp. 1136–1139.
27. C. Busso and S. Narayanan, "Interrelation between Speech and Facial Gestures in Emotional Utterances: A Single Subject Study," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, 2007, pp. 2331–2347.
28. Y. Wang and L. Guan, "Recognizing Human Emotional State from Audiovisual Signals," *IEEE Trans. Multimedia*, vol. 10, no. 5, 2008, pp. 659–668.
29. B. Schuller et al., eds., *Proc. Audio/Visual Emotion Challenge and Workshop*, Social Signal Processing Network (SSPNET), 2011.
30. E. Cambria et al., *Sentic Avatar: Multimodal Affective Conversational Agent with Common Sense*, Springer, 2011.

and segmentation are pretty reliable, but also time consuming. Alternatives for performing the transcription step automatically include the use of automatic speech recognition technologies, such as Google Voice or Adobe Translator, or crowdsourcing techniques, such as Amazon Mechanical Turk. Later, we present an English dataset, which was efficiently transcribed using this crowdsourcing approach.

Sentiment Annotations

Because our goal is to automatically find the sentiment expressed in the video clip, we decided to perform our annotation task at the video-sequence level. This is an important step in creating the dataset, and we were particularly careful while describing the task. We asked 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 assigned one of 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 percent, with a κ of 0.84, which indicates strong agreement. To determine the final gold-standard

label, all of the annotation disagreements were resolved through discussion. The final dataset consists of 105 video clips, of which 47 are labeled as positive, 54 as negative, and 4 as neutral. The dataset's baseline is 51 percent, which corresponds to the accuracy obtained if all of the videos are assigned with the most frequent polarity label in the dataset.

Multimodal Sentiment Analysis

The greatest advantage of analyzing video opinions as compared to text-only opinions is that we can use additional cues. In textual opinions, the only available source of information consists of the words in the opinion and the dependencies among them, which might sometimes prove insufficient to convey the consumer's exact sentiment. 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 speaker's tone, whereas visual data can provide information regarding the speaker's emotional state. Thus, our hypothesis is that a combination of text and video data can help create a better analysis model. We specifically focus on three main types of features covering the three modalities.

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 stop words—occurring in the transcriptions of the training set. We then remove those words that have a frequency below 10 (a 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 (SVMs), they've been shown to lead to state-of-the-art performance.[9,10]

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 (30 Hz), with a sliding window of 50 milliseconds (ms). We used the open source software OpenEAR[15] 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 every experiment and speaker.

For each video in our dataset, we defined four summary features: pause duration, pitch, intensity, and loudness.

Pause duration. Given the audio frames extracted from the entire video, this determines how many audio samples are identified as silent. This audio feature is then normalized by the number of audio samples in the video. This feature can be interpreted as the percentage of time in which the speaker was silent.

Pitch. This computes the standard deviation of the video's pitch level. This measure represents the variation of voice intonation during the entire video.

Intensity. This measures the sound power of the spoken utterances in the video. We compute the average voice intensity over the whole video.

Loudness. This determines the perceived strength of the voice factored by the ear's sensitivity. We compute the average loudness measure over the entire video.

Visual Features

The visual features are automatically extracted from the video sequences. Because only one person is present in each video clip, and most of the time that person is 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, extracts the facial features, and extrapolates some basic facial expressions as well as eye gaze direction. The main facial expression that it recognizes is a 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 frame rate: 30 Hz.

An important aspect when generating visual features is the video quality, and correspondingly, the visual processing quality that can be automatically performed on the video. OKAO provides a confidence level for each processed frame in the range 0–1,000. We discounted all the frames with a confidence level below 700, and we also removed any videos where more than 30 percent of the frames had a confidence level below 700.

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

- *Smile duration.* Given all the frames in a video, this feature determines how many frames are identified as a smile. In our experiments, we use three different variants of this feature

Table 1. Automatic sentiment classification performance for seven different models on our Spanish multimodal opinion dataset.

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**

with different thresholds: 50 and 75 frames.

- *Look-away duration*. Given all the frames in a video, this feature measures the number of frames where the speaker is looking at the camera. The horizontal and vertical angular thresholds were experimentally set to 10 degrees.

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

Experiments

We ran our main experiments on the Spanish multimodal opinion dataset. From this dataset, we remove any videos that had low visual-processing performance (for example, if the number of frames correctly processed by OKAO was below 70 percent), and further remove videos labeled as neutral (thereby keeping only positive and negative videos). This left us with an experimental dataset of 76 videos, consisting of 39 positive and 37 negative videos, for which we extracted linguistic, audio, and visual features.

The multimodal fusion was performed using the early fusion technique, where all the linguistic, audio, and visual features were concatenated into a common feature vector, thus resulting in one vector for each video in the dataset. For classification, we used SVMs 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 a maximum margin.[16] We used the Weka machine-learning toolkit. For each experiment, a 10-fold cross validation was run on the entire dataset.

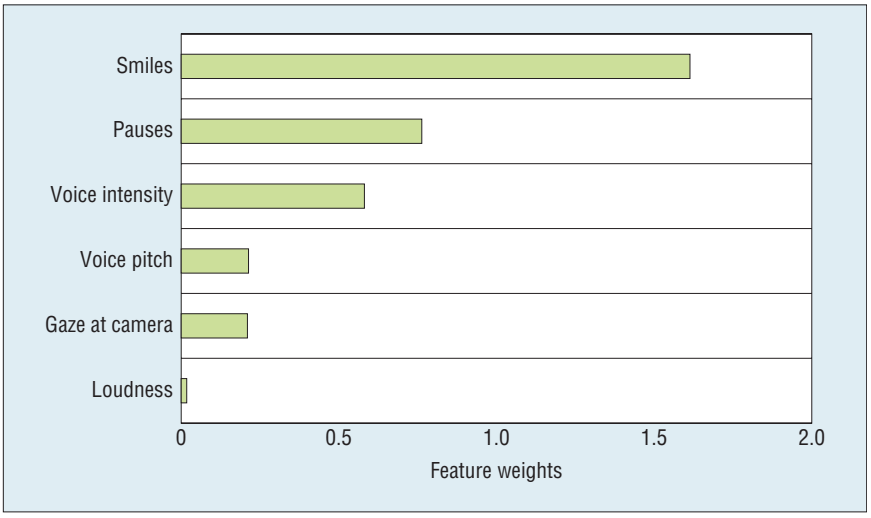

Figure 3. Visual and audio feature weights. This graph shows the relative importance of the support vector machine (SVM) weights associated with each audio-visual feature.

Results and Discussions

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 then the audio classifier.

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 Figure 3 shows. Perhaps unsurprisingly, the smile is the most predictive feature, followed by the number of pauses and voice intensity. Voice pitch, gaze at camera, and loudness also contribute 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. 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 higher voice intensity and speak at a higher rate (so that they pause less), unlike the speakers of a positive review who tend to speak at a slower pace.

Multimodal Sentiment Analysis on English Videos

As a final experiment, to determine the multimodal sentiment analysis method's portability to a different dataset, we compile a second dataset consisting of English video reviews. We collect cellular phone reviews from ExpoTV (www.expotv.com), which is a public website that provides consumer-generated videos. Through this platform, users provide unbiased video opinions of products organized in various categories.

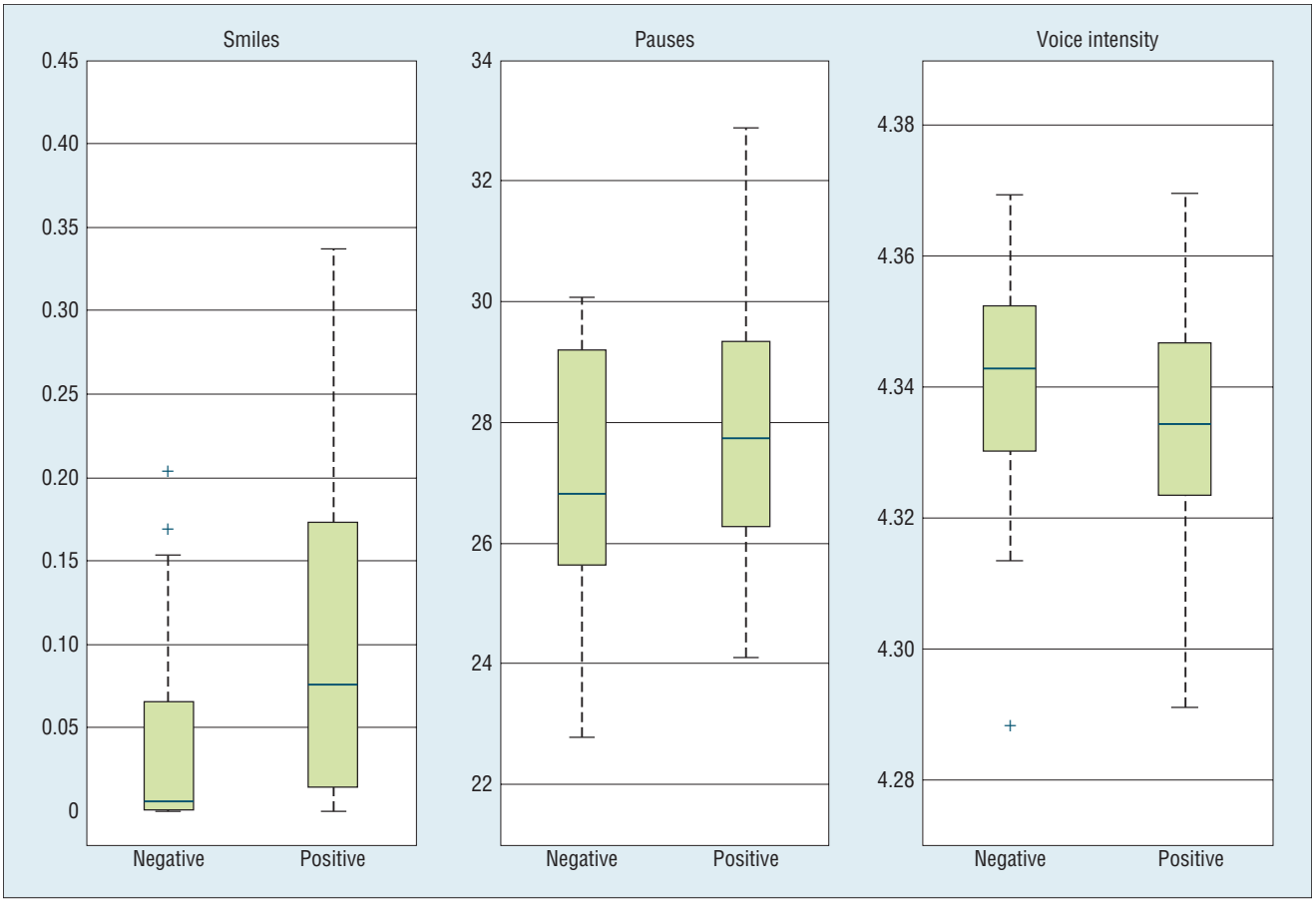

Figure 4. Average values of several multimodal features when clustered per sentiment label. An increased number of smiles and an increased number of pauses are characteristic of positive videos, whereas higher voice intensity and fewer pauses are more typical of negative videos.

Table 2. Multimodal sentiment analysis on an English dataset.

Modality	Accuracy (%)
Text only	64.94
Visual only	61.04
Audio only	46.75
Text-audio-visual	**64.86**

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 often showing the product they review to the camera, thus covering their face. Because 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-second segment in which people express their opinion. To obtain the transcriptions, this time we used crowdsourcing via Amazon Mechanical Turk. To ensure quality, one of us personally verified the transcriptions collected from the Amazon service. ExpoTV users provide a star rating to the product they're reviewing (one to five stars). Thus, for the sentiment annotations, we used this rating 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 and 17 negative reviews.

Table 2 shows the results obtained on the English dataset. As with our results on the Spanish dataset (see Table 1), the joint use of all three modalities brings significant improvements over models that use only one modality at a time. Interestingly, here again, the audio model is the weakest model, which suggests audio feature engineering as a possible avenue for future work.

As the quantity of online opinion videos increases, current sentiment analysis techniques need to be extended so that they can better handle the presence of multiple modalities. Here, we presented our initial efforts towards automatic identification of expressed sentiment in short opinion segments. Our results are promising and pave the way to a new line of research for multimodal sentiment analysis.

In the short term, our future work includes an exploration of our proposed approach for full-length videos, which might contain a mixture of positive, negative, and neutral segments. This poses additional challenges, such as segmenting the video content at the appropriate utterance level, working with smaller data units for the visual and acoustic analysis, and improving the data fusion process. In the longer term, additional research is needed to explore datasets covering other domains and languages, giving us a better understanding of cultural differences.

The datasets introduced in this article are available upon request. ▪

Acknowledgments

This material is based in part upon work supported by US National Science Foundation awards 0917170 and 1118018. 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," *Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, ACM, 2004, pp. 168–177.
2. C. Alm, D. Roth, and R. Sproat, "Emotions from Text: Machine Learning for Text-Based Emotion Prediction," *Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing*, Assoc. for Computational Linguistics (ACL), 2005, pp. 579–589.
3. L. Lloyd, D. Kechagias, and S. Skiena, "Lydia: A System for Large-Scale News Analysis," *String Processing and Information Retrieval*, LNCS 3772, 2005, pp. 161–166.
4. P. Carvalho et al., "Liars and Saviors in a Sentiment Annotated Corpus of Comments to Political Debates," *Proc. Assoc. for Computational Linguistics*, ACL, 2011, pp. 564–568.
5. H. Yu and V. Hatzivassiloglou, "Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences," *Proc. Conf. Empirical Methods in Natural Language Processing*, ACL, 2003, pp. 129–136.
6. G. Carenini, R. Ng, and X. Zhou, "Summarizing Emails with Conversational Cohesion and Subjectivity," *Proc. Assoc. for Computational Linguistics: Human Language Technologies*, ACL, 2008, pp. 353–361.
7. J. Wiebe and E. Riloff, "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts," *Proc. 6th Int'l Conf. Intelligent Text Processing and Computational Linguistics*, Springer, 2005, pp. 486–497.
8. A. Esuli and F. Sebastiani, "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining," *Proc. 5th Conf. Language Resources and Evaluation*, Springer, 2006, pp. 417–422.
9. B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *Proc. 42nd Meeting of the Assoc. Computational Linguistics*, ACL, 2004, pp. 271–278.
10. A. Maas et al., "Learning Word Vectors for Sentiment Analysis," *Proc. Assoc. Computational Linguistics*, ACL, 2011, pp. 142–150.
11. D. Chen and R. Mooney, "Panning for Gold: Finding Relevant Semantic Content for Grounded Language Learning," *Proc. Symp. Machine Learning in Speech and Language Processing*, 2011; www.cs.utexas.edu/~ml/papers/chen.mlslp11.pdf.
12. C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual Sentiment and Subjectivity," *Multilingual Natural Language Processing*, Prentice Hall, 2011.
13. T. Plotz and G.A. Fink, "Markov Models for Offline Handwriting Recognition: A Survey," *Int'l J. Document Analysis and Recognition*, vol. 12, no. 4, 2009.
14. J. Arguello and C. Rose, "Topic Segmentation of Dialogue," *Proc. HLT-NAACL Workshop on Analyzing Conversations in Text and Speech*, ACL, 2009, pp. 42–49.
15. F. Eyben, M. Wollmer, and B. Schuller, "OpenEAR—Introducing the Munich Open Source Emotion and Affect Recognition Toolkit," *Proc. Affective Computing and Intelligent Interaction and Workshops*, IEEE, 2009; doi:10.1109/ACII2009.5349350.
16. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

THE AUTHORS

Verónica Pérez Rosas is a doctoral student in the Department of Computer Science and Engineering at the University of North Texas. Her research interests include natural language processing, machine learning, and intelligent optimization. Pérez Rosas has an MS in computer science from the Instituto Tecnólogico de Ciudad Madero. Contact her at veronicaperezrosas@my.unt.edu.

Rada Mihalcea is an associate professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. 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. Mihalcea has a PhD in computer science and engineering from Southern Methodist University and a PhD in linguistics from Oxford University. She is the recipient of a National Science Foundation Career award and a Presidential Early Career Award for Scientists and Engineers. Contact her at mihalcea@umich.edu.

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. His research interests focus on the computational study of nonverbal social communication, a multidisciplinary research topic that overlays the fields of multimodal interaction, computer vision, machine learning, social psychology, and artificial intelligence. Morency has a PhD in computer science and artificial intelligence from MIT. In 2008, *IEEE Intelligent Systems* selected him as one of the "10 to watch" for the future of AI research. Contact him at morency@ict.usc.edu.