Utterance-Level Multimodal Sentiment Analysis

Verónica Pérez-Rosas and Rada Mihalcea
Computer Science and Engineering
University of North Texas
veronicaperezrosas@my.unt.edu, rada@cs.unt.edu

Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
morency@ict.usc.edu

Abstract

During real-life interactions, people are naturally gesturing and modulating their voice to emphasize specific points or to express their emotions. With the recent growth of social websites such as YouTube, Facebook, and Amazon, video reviews are emerging as a new source of multimodal and natural opinions that has been left almost untapped by automatic opinion analysis techniques. This paper presents a method for multimodal sentiment classification, which can identify the sentiment expressed in utterance-level visual datastreams. Using a new multimodal dataset consisting of sentiment annotated utterances extracted from video reviews, we show that multimodal sentiment analysis can be effectively performed, and that the joint use of visual, acoustic, and linguistic modalities can lead to error rate reductions of up to 10.5% as compared to the best performing individual modality.

1 Introduction

Video reviews represent a growing source of consumer information that gained increasing interest from companies, researchers, and consumers. Popular web platforms such as YouTube, Amazon, Facebook, and ExpoTV have reported a significant increase in the number of consumer reviews in video format over the past five years. Compared to traditional text reviews, video reviews provide a more natural experience as they allow the viewer to better sense the reviewer's emotions, beliefs, and intentions through richer channels such as intonations, facial expressions, and body language.

Much of the work to date on opinion analysis has focused on textual data, and a number of resources have been created including lexicons (Wiebe and Riloff, 2005; Esuli and Sebastiani, 2006) or large annotated datasets (Maas et al., 2011). 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. This has motivated researchers to start exploring multimodal clues for the detection of sentiment and emotions in video content (Morency et al., 2011; Wagner et al., 2011).

In this paper, we explore the addition of speech and visual modalities to text analysis in order to identify the sentiment expressed in video reviews. Given the non homogeneous nature of full-video reviews, which typically include a mixture of positive, negative, and neutral statements, we decided to perform our experiments and analyses at the utterance level. This is in line with earlier work on text-based sentiment analysis, where it has been observed that full-document reviews often contain both positive and negative comments, which led to a number of methods addressing opinion analysis at sentence level. Our results show that relying on the joint use of linguistic, acoustic, and visual modalities allows us to better sense the sentiment being expressed as compared to the use of only one modality at a time.

Another important aspect of this paper is the introduction of a new multimodal opinion database annotated at the utterance level which is, to our knowledge, the first of its kind. In our work, this dataset enabled a wide range of multimodal sentiment analysis experiments, addressing the relative importance of modalities and individual features.

The following section presents related work in text-based sentiment analysis and audio-visual emotion recognition. Section 3 describes our new multimodal datasets with utterance-level sentiment annotations. Section 4 presents our multimodal sen-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 973–982,
Sofia, Bulgaria, August 4-9 2013.

timent analysis approach, including details about our linguistic, acoustic, and visual features. Our experiments and results on multimodal sentiment classification are presented in Section 5, with a detailed discussion and analysis in Section 6.

2 Related Work

In this section we provide a brief overview of related work in text-based sentiment analysis, as well as audio-visual emotion analysis.

2.1 Text-based Subjectivity and Sentiment Analysis

The techniques developed so far for subjectivity and sentiment analysis have focused primarily on the processing of text, and 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. These tools and resources have been already used in a large number of applications, including expressive text-to-speech synthesis (Alm et al., 2005), tracking sentiment timelines in on-line forums and news (Balog et al., 2006), analysis of political debates (Carvalho et al., 2011), question answering (Oh et al., 2012), conversation summarization (Carenini et al., 2008), and citation sentiment detection (Athar and Teufel, 2012).

One of the first lexicons used in sentiment analysis is the General Inquirer (Stone, 1968). Since then, many methods have been developed to automatically identify opinion words and their polarity (Hatzivassiloglou and McKeown, 1997; Turney, 2002; Hu and Liu, 2004; Taboada et al., 2011), as well as n-gram and more linguistically complex phrases (Yang and Cardie, 2012).

For data-driven methods, one of the most widely used datasets is the MPQA corpus (Wiebe et al., 2005), which is a collection of news articles manually annotated for opinions. Other datasets are also available, including two polarity datasets consisting of movie reviews (Pang and Lee, 2004; Maas et al., 2011), and a collection of newspaper headlines annotated for polarity (Strapparava and Mihalcea, 2007).

While difficult problems such as cross-domain (Blitzer et al., 2007; Li et al., 2012) or cross-language (Mihalcea et al., 2007; Wan, 2009; Meng et al., 2012) portability have been addressed, not much has been done in terms of extending the applicability of sentiment analysis to other modalities, such as speech or facial expressions.

The only exceptions that we are aware of are the findings reported in (Somasundaran et al., 2006; Raaijmakers et al., 2008; Mairesse et al., 2012; Metze et al., 2009), where speech and text have been analyzed jointly for the purpose of subjectivity or sentiment identification, without, however, addressing other modalities such as visual cues; and the work reported in (Morency et al., 2011; Perez-Rosas et al., 2013), where multimodal cues have been used for the analysis of sentiment in product reviews, but where the analysis was done at the much coarser level of full videos rather than individual utterances as we do in our work.

2.2 Audio-Visual Emotion Analysis.

Also related to our work is the research done on emotion analysis. Emotion analysis of speech signals aims to identify the emotional or physical states of a person by analyzing his or her voice (Ververidis and Kotropoulos, 2006). Proposed methods for emotion recognition from speech focus both on what is being said and how is being said, and rely mainly on the analysis of the speech signal by sampling the content at utterance or frame level (Bitouk et al., 2010). Several researchers used prosody (e.g., pitch, speaking rate, Mel frequency coefficients) for speech-based emotion recognition (Polzin and Waibel, 1996; Tato et al., 2002; Ayadi et al., 2011).

There are also studies that analyzed the visual cues, such as facial expressions and body movements (Calder et al., 2001; Rosenblum et al., 1996; Essa and Pentland, 1997). Facial expressions are among the most powerful and natural means for human beings to communicate their emotions and intentions (Tian et al., 2001). Emotions can be also expressed unconsciously, through subtle movements of facial muscles such as smiling or eyebrow raising, often measured and described using the Facial Action Coding System (FACS) (Ekman et al., 2002).

De Silva et. al. (De Silva et al., 1997) and Chen et. al. (Chen et al., 1998) presented one of the early works that integrate both acoustic and visual information for emotion recognition. In addition to work that considered individual modalities, there is also a growing body of work concerned with multimodal emotion analysis (Silva et al., 1997; Sebe et al., 2006; Zhihong et al., 2009; Wollmer et al., 2010).

Utterance transcription	Label
En este color, creo que era el color frambuesa. In this color, I think it was raspberry	neu
Pinta hermosisimo. It looks beautiful.	pos
Sinceramente, con respecto a lo que pinta y a que son hidratante, si son muy hidratantes. Honestly, talking about how they looks and hydrates, yes they are very hydrant.	pos
Pero el problema de estos labiales es que cuando uno se los aplica, te dejan un gusto asqueroso en la boca. But the problem with those lipsticks is that when you apply them, they leave a very nasty taste	neg
Sinceramente, es no es que sea el olor sino que es mas bien el gusto. Honestly, is not the smell, it is the taste.	neg

Table 1: Sample utterance-level annotations. The labels used are: pos(itive), neg(ative), neu(tral).

More recently, two challenges have been organized focusing on the recognition of emotions using audio and visual cues (Schuller et al., 2011a; Schuller et al., 2011b), which included subchallenges on audio-only, video-only, and audio-video, and drew the participation of many teams from around the world. Note however that most of the previous work on audio-visual emotion analysis has focused exclusively on the audio and video modalities, and did not consider textual features, as we do in our work.

3 MOUD: Multimodal Opinion Utterances Dataset

For our experiments, we created a dataset of utterances (named MOUD) containing product opinions expressed in Spanish.[1] We chose to work with Spanish because it is a widely used language, and it is the native language of the main author of this paper.

We started by collecting a set of videos from the social media web site YouTube, using several keywords likely to lead to a product review or recommendation. Starting with the YouTube search page, videos were found using the following keywords: *mis products favoritos* (my favorite products), *products que no recomiendo* (non recommended products), *mis perfumes favoritos* (my favorite perfumes), *peliculas recomendadas* (recommended movies), *peliculas que no recomiendo* (non recommended movies) and *libros recomendados* (recommended books), *libros que no recomiendo* (non recommended books). Notice that the keywords are not targeted at a specific product type; rather, we used a variety of product names, so that the dataset has some degree of generality within the broad domain of product reviews.

[1]Publicly available from the authors webpage.

Among all the videos returned by the YouTube search, we selected only videos that respected the following guidelines: the speaker should be in front of the camera; her face should be clearly visible, with a minimum amount of face occlusion during the recording; there should not be any background music or animation. The final video set includes 80 videos randomly selected from the videos retrieved from YouTube that also met the guidelines above. The dataset includes 15 male and 65 female speakers, with their age approximately ranging from 20 to 60 years.

All the videos were first pre-processed to eliminate introductory titles and advertisements. Since the reviewers often switched topics when expressing their opinions, we manually selected a 30 seconds opinion segment from each video to avoid having multiple topics in a single review.

3.1 Segmentation and Transcription

All the video clips were manually processed to transcribe the verbal statements and also to extract the start and end time of each utterance. Since the reviewers utter expressive sentences that are naturally segmented by speech pauses, we decided to use these pauses (>0.5seconds) to identify the beginning and the end of each utterance. The transcription and segmentation were performed using the Transcriber software.

Each video was segmented into an average of six utterances, resulting in a final dataset of 498 utterances. Each utterance is linked to the corresponding audio and video stream, as well as its manual transcription. The utterances have an average duration of 5 seconds, with a standard deviation of 1.2 seconds.

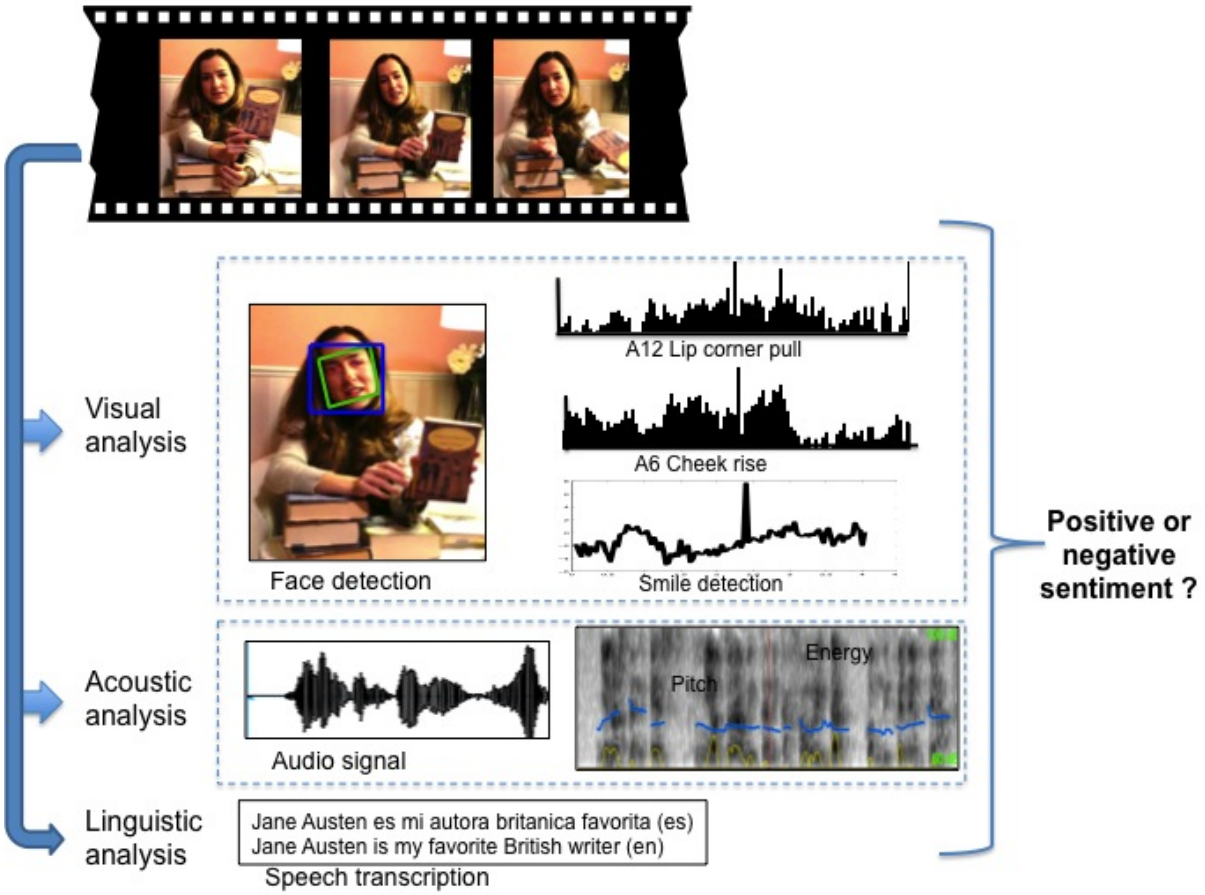

Figure 1: Multimodal feature extraction

3.2 Sentiment Annotation

To enable the use of this dataset for sentiment detection, we performed sentiment annotations at utterance level. Annotations were done using Elan,[2] which is a widely used tool for the annotation of video and audio resources. Two annotators independently labeled each utterance as positive, negative, or neutral. The annotation was done after seeing the video corresponding to an utterance (along with the corresponding audio source). The transcription of the utterance was also made available. Thus, the annotation process included all three modalities: visual, acoustic, and linguistic. The annotators were allowed to watch the video segment and their corresponding transcription as many times as needed.

The inter-annotator agreement was measured at 88%, with a Kappa of 0.81, which represents good agreement. All the disagreements were reconciled through discussions.

Table 1 shows the five utterances obtained from a video in our dataset, along with their corresponding sentiment annotations. As this example illustrates, a video can contain a mix of positive, negative, and neutral utterances. Note also that sentiment is not always explicit in the text: for example, the last utterance "Honestly, it is not the smell, it is the taste" has an implicit reference to the "nasty taste" expressed in the previous utterance, and thus it was also labeled as negative by both annotators.

4 Multimodal Sentiment Analysis

The main advantage that comes with the analysis of video opinions, as compared to their textual counterparts, is the availability of visual and speech cues. In textual opinions, the only source of information consists of words and their dependencies, which may sometime prove insufficient to convey the exact sentiment of the user. Instead, video opinions naturally contain multiple modalities, consisting of visual, acoustic, and linguistic datastreams. We hypothesize that the simultaneous use of these three modalities will help create a better opinion analysis model.

[2] http://tla.mpi.nl/tools/tla-tools/elan/

4.1 Feature Extraction

This section describes the process of automatically extracting linguistic, acoustic and visual features from the video reviews. First, we obtain the stream corresponding to each modality, followed by the extraction of a representative set of features for each modality, as described in the following subsections. These features are then used as cues to build a classifier of positive or negative sentiment. Figure 1 illustrates this process.

4.1.1 Linguistic Features

We use a bag-of-words representation of the video transcriptions of each utterance to derive unigram counts, which are then used as linguistic 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 utterance 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 (SVM) have been shown to lead to state-of-the-art performance (Maas et al., 2011).

4.1.2 Acoustic Features

Acoustic features are automatically extracted from the speech signal of each utterance. We used the open source software OpenEAR (Schuller, 2009) to automatically compute a set of acoustic features. We include prosody, energy, voicing probabilities, spectrum, and cepstral features.

- Prosody features. These include intensity, loudness, and pitch that describe the speech signal in terms of amplitude and frequency.
- Energy features. These features describe the human loudness perception.
- Voice probabilities. These are probabilities that represent an estimate of the percentage of voiced and unvoiced energy in the speech.
- Spectral features. The spectral features are based on the characteristics of the human ear, which uses a nonlinear frequency unit to simulate the human auditory system. These features describe the speech formants, which model spoken content and represent speaker characteristics.
- Cepstral features. These features emphasize changes or periodicity in the spectrum features measured by frequencies; we model them using 12 Mel-frequency cepstral coefficients that are calculated based on the Fourier transform of a speech frame.

Overall, we have a set of 28 acoustic features. During the feature extraction, we use a frame sampling of 25ms. Speaker normalization is performed using z-standardization. The voice intensity is thresholded to identify samples with and without speech, with the same threshold being used for all the experiments and all the speakers. The features are averaged over all the frames in an utterance, to obtain one feature vector for each utterance.

4.1.3 Facial Features

Facial expressions can provide important clues for affect recognition, which we use to complement the linguistic and acoustic features extracted from the speech stream.

The most widely used system for measuring and describing facial behaviors is the Facial Action Coding System (FACS), which allows for the description of face muscle activities through the use of a set of Action Units (AUs). According with (Ekman, 1993), there are 64 AUs that involve the upper and lower face, including several face positions and movements.[3] AUs can occur either by themselves or in combination, and can be used to identify a variety of emotions. While AUs are frequently annotated by certified human annotators, automatic tools are also available. In our work, we use the Computer Expression Recognition Toolbox (CERT) (Littlewort et al., 2011), which allows us to automatically extract the following visual features:

- Smile and head pose estimates. The smile feature is an estimate for smiles. Head pose detection consists of three-dimensional estimates of the head orientation, i.e., yaw, pitch, and roll. These features provide information about changes in smiles and face positions while uttering positive and negative opinions.
- Facial AUs. These features are the raw estimates for 30 facial AUs related to muscle movements for the eyes, eyebrows, nose, lips,

[3] http://www.cs.cmu.edu/afs/cs/project/face/www/facs.htm

and chin. They provide detailed information about facial behaviors from which we expect to find differences between positive and negative states.

- Eight basic emotions. These are estimates for the following emotions: anger, contempt, disgust, fear, joy, sad, surprise, and neutral. These features describe the presence of two or more AUs that define a specific emotion. For example, the unit A12 describes the pulling of lip corners movement, which usually suggests a smile but when associated with a check raiser movement (unit A6), represents a marker for the emotion of happiness.

We extract a total of 40 visual features, each of them obtained at frame level. Since only one person is present in each video clip, most of the time facing the camera, the facial tracking was successfully applied for most of our data. For the analysis, we use a sampling rate of 30 frames per second. The features extracted for each utterance are averaged over all the valid frames, which are automatically identified using the output of CERT.[4] Segments with more than 60% of invalid frames are simply discarded.

5 Experiments and Results

We run our sentiment classification experiments on the MOUD dataset introduced earlier. From the dataset, we remove utterances labeled as neutral, thus keeping only the positive and negative utterances with valid visual features. The removal of neutral utterances is done for two main reasons. First, the number of neutral utterances in the dataset is rather small. Second, previous work in subjectivity and sentiment analysis has demonstrated that a layered approach (where neutral statements are first separated from opinion statements followed by a separation between positive and negative statements) works better than a single three-way classification. After this process, we are left with an experimental dataset of 412 utterances, 182 of which are labeled as positive, and 231 are labeled as negative.

From each utterance, we extract the linguistic, acoustic, and visual features described above, which are then combined using the early fusion (or feature-level fusion) approach (Hall and Llinas, 1997; Atrey et al., 2010). In this approach, the features collected from all the multimodal streams are combined into a single feature vector, thus resulting in one vector for each utterance in the dataset which is used to make a decision about the sentiment orientation of the utterance.

[4]There is a small number of frames that CERT could not process, mostly due to the brief occlusions that occur when the speaker is showing the product she is reviewing.

Modality	Accuracy
Baseline	55.93%
One modality at a time	
Linguistic	70.94%
Acoustic	64.85%
Visual	67.31%
Two modalities at a time	
Linguistic + Acoustic	72.88%
Linguistic + Visual	72.39%
Acoustic + Visual	68.86%
Three modalities at a time	
Linguistic+Acoustic+Visual	74.09%

Table 2: Utterance-level sentiment classification with linguistic, acoustic, and visual features.

We run several comparative experiments, using one, two, and three modalities at a time. We use the entire set of 412 utterances and run ten fold cross validations using an SVM classifier, as implemented in the Weka toolkit.[5] In line with previous work on emotion recognition in speech (Haq and Jackson, 2009; Anagnostopoulos and Vovoli, 2010) where utterances are selected in a speaker dependent manner (i.e., utterances from the same speaker are included in both training and test), as well as work on sentence-level opinion classification where document boundaries are not considered in the split performed between the training and test sets (Wilson et al., 2004; Wiegand and Klakow, 2009), the training/test split for each fold is performed at utterance level regardless of the video they belong to.

Table 2 shows the results of the utterance-level sentiment classification experiments. The baseline is obtained using the ZeroR classifier, which assigns the most frequent label by default, averaged over the ten folds.

6 Discussion

The experimental results show that sentiment classification can be effectively performed on multimodal datastreams. Moreover, the integration of

[5]http://www.cs.waikato.ac.nz/ml/weka/

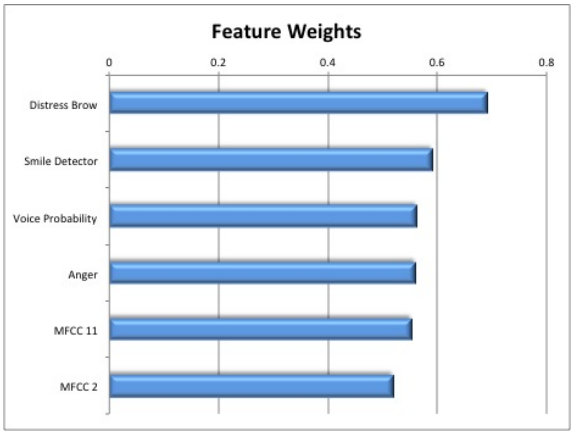

Figure 2: Visual and acoustic feature weights. This graph shows the relative importance of the information gain weights associated with the top most informative acoustic-visual features.

visual, acoustic, and linguistic features can improve significantly over the use of one modality at a time, with incremental improvements observed for each added modality.

Among the individual classifiers, the linguistic classifier appears to be the most accurate, followed by the classifier that relies on visual clues, and by the audio classifier. Compared to the best individual classifier, the relative error rate reduction obtained with the tri-modal classifier is 10.5%. The results obtained with this multimodal utterance classifier are found to be significantly better than the best individual results (obtained with the text modality), with significance being tested with a t-test (p=0.05).

Feature analysis.
To determine the role played by each of the visual and acoustic features, we compare the feature weights assigned by the learning algorithm, as shown in Figure 2. Interestingly, a distressed brow is the strongest indicator of sentiment, followed, this time not surprisingly, by the smile feature. Other informative features for sentiment classification are the voice probability, representing the energy in speech, the combined visual features that represent an angry face, and two of the cepstral coefficients.

To reach a better understanding of the relation between features, we also calculate the Pearson correlation between the visual and acoustic features. Table 3 shows a subset of these correlation figures. As we expected, correlations between features of the same type are higher. For example, the correlation between features AU6 and AU12 or the correlation between intensity and loudness is higher than the correlation between AU6 and intensity. Nonetheless, we still find some significant correlations between features of different types, for instance AU12 and AU45 which are both significantly correlated with the intensity and loudness features. This give us confidence about using them for further analysis.

Video-level sentiment analysis.
To understand the role played by the size of the video-segments considered in the sentiment classification experiments, as well as the potential effect of a speaker-independence assumption, we also run a set of experiments where we use full videos for the classification.

In these experiments, once again the sentiment annotation is done by two independent annotators, using the same protocol as in the utterance-based annotations. Videos that were ambivalent about the general sentiment were either labeled as neutral (and thus removed from the experiments), or labeled with the dominant sentiment. The inter-annotator agreement for this annotation was measured at 96.1%. As before, the linguistic, acoustic, and visual features are averaged over the entire video, and we use an SVM classifier in ten-fold cross validation experiments.

Table 4 shows the results obtained in these video-level experiments. While the combination of modalities still helps, the improvement is smaller than the one obtained during the utterance-level classification. Specifically, the combined effect of acoustic and visual features improves significantly over the individual modalities. However, the combination of linguistic features with other modalities does not lead to clear improvements. This may be due to the smaller number of feature vectors used in the experiments (only 80, as compared to the 412 used in the previous setup). Another possible reason is the fact that the acoustic and visual modalities are significantly weaker than the linguistic modality, most likely due to the fact that the feature vectors are now speaker-independent, which makes it harder to improve over the linguistic modality alone.

7 Conclusions

In this paper, we presented a multimodal approach for utterance-level sentiment classification. We introduced a new multimodal dataset consisting

	AU6	AU12	AU45	AUs 1,1+4	Pitch	Voice probability	Intensity	Loudness
AU6	1.00	0.46*	-0.03	-0.05	0.06	-0.14*	-0.04	-0.02
AU12		1.00	-0.23*	-0.33*	0.04	0.05	0.15*	0.16*
AU45			1.00	0.05	-0.05	-0.11*	-.163*	0.16*
AUs 1,1+4				1.00	-0.11*	-0.16*	0.06	0.07
Pitch					1.00	-0.04	-0.01	-0.08
Voice probability						1.00	0.19*	0.38*
Intensity							1.00	0.85*
Loudness								1.00

Table 3: Correlations between several visual and acoustic features. Visual features: AU6 Cheek raise, AU12 Lip corner pull, AU45 Blink eye and closure, AU1,1+4 Distress brow. Acoustic features: Pitch, Voice probability, Intensity, Energy. *Correlation is significant at the 0.05 level (1-tailed)

.

Modality	Accuracy
Baseline	55.93%
One modality at a time	
Linguistic	73.33%
Acoustic	53.33%
Visual	50.66%
Two modalities at a time	
Linguistic + Acoustic	72.00%
Linguistic + Visual	74.66%
Acoustic + Visual	61.33%
Three modalities at a time	
Linguistic+Acoustic+Visual	74.66%

Table 4: Video-level sentiment classification with linguistic, acoustic, and visual features.

of sentiment annotated utterances extracted from video reviews, where each utterance is associated with a video, acoustic, and linguistic datastream. Our experiments show that sentiment annotation of utterance-level visual datastreams can be effectively performed, and that the use of multiple modalities can lead to error rate reductions of up to 10.5% as compared to the use of one modality at a time. In future work, we plan to explore alternative multimodal fusion methods, such as decision-level and meta-level fusion, to improve the integration of the visual, acoustic, and linguistic modalities.

Acknowledgments

We would like to thank Alberto Castro for his help with the sentiment annotations. This material is based in part upon work supported by National Science Foundation awards #0917170 and #1118018, by DARPA-BAA-12-47 DEFT grant #12475008, and by a grant from U.S. RDECOM. 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, the Defense Advanced Research Projects Agency, or the U.S. Army Research, Development, and Engineering Command.

References

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

C. Anagnostopoulos and E. Vovoli. 2010. Sound processing features for speaker-dependent and phrase-independent emotion recognition in berlin database. In *Information Systems Development*, pages 413–421. Springer.

A. Athar and S. Teufel. 2012. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, June.

P. K. Atrey, M. A. Hossain, A. El Saddik, and M. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16.

M. El Ayadi, M. Kamel, and F. Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572 – 587.

K. Balog, G. Mishne, and M. de Rijke. 2006. 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)*.

Dmitri Bitouk, Ragini Verma, and Ani Nenkova. 2010. Class-level spectral features for emotion recognition. *Speech Commun.*, 52(7-8):613–625, July.

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Association for Computational Linguistics*.

A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu. 2001. A principal component analysis of facial expressions. *Vision research*, 41(9):1179–1208, April.

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

P. Carvalho, L. Sarmento, J. Teixeira, and M. Silva. 2011. 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.

L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu. 1998. Multimodal human emotion/expression recognition. In *Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, pages 366–, Washington, DC, USA. IEEE Computer Society.

L C De Silva, T Miyasato, and R Nakatsu, 1997. *Facial emotion recognition using multi-modal information*, volume 1, page 397401. IEEE Signal Processing Society.

P. Ekman, W. Friesen, and J. Hager. 2002. Facial action coding system.

P. Ekman. 1993. Facial expression of emotion. *American Psychologist*, 48:384–392.

I.A. Essa and A.P. Pentland. 1997. Coding, analysis, interpretation, and recognition of facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):757 –763, jul.

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

D.L. Hall and J. Llinas. 1997. An introduction to multisensor fusion. *IEEE Special Issue on Data Fusion*, 85(1).

S. Haq and P. Jackson. 2009. Speaker-dependent audio-visual emotion recognition. In *International Conference on Audio-Visual Speech Processing*.

V. Hatzivassiloglou and K. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181.

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

F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu. 2012. Cross-domain co-extraction of sentiment and topic lexicons. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea.

G. Littlewort, J. Whitehill, Tingfan Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. 2011. The computer expression recognition toolbox (cert). In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298 –305, march.

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

F. Mairesse, J. Polifroni, and G. Di Fabbrizio. 2012. Can prosody inform sentiment analysis? experiments on short spoken reviews. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5093 –5096, march.

X. Meng, F. Wei, X. Liu, M. Zhou, G. Xu, and H. Wang. 2012. Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea.

F. Metze, T. Polzehl, and M. Wagner. 2009. Fusion of acoustic and linguistic features for emotion detection. In *Semantic Computing, 2009. ICSC '09. IEEE International Conference on*, pages 153 –160, sept.

R. Mihalcea, C. Banea, and J. Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the Association for Computational Linguistics*, Prague, Czech Republic.

L.P. Morency, R. Mihalcea, and P. Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the International Conference on Multimodal Computing*, Alicante, Spain.

J. Oh, K. Torisawa, C. Hashimoto, T. Kawada, S. De Saeger, J. Kazama, and Y. Wang. 2012. Why question answering using sentiment analysis and word classes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea.

B. Pang and L. Lee. 2004. 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.

V. Perez-Rosas, R. Mihalcea, and L.-P. Morency. 2013. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*.

T. Polzin and A. Waibel. 1996. Recognizing emotions in speech. In *In ICSLP*.

S. Raaijmakers, K. Truong, and T. Wilson. 2008. Multimodal subjectivity analysis of multiparty conversation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 466–474, Honolulu, Hawaii.

M. Rosenblum, Y. Yacoob, and L.S. Davis. 1996. Human expression recognition from motion using a radial basis function network architecture. *Neural Networks, IEEE Transactions on*, 7(5):1121 –1138, sep.

B. Schuller, M. Valstar, R. Cowie, and M. Pantic, editors. 2011a. *Audio/Visual Emotion Challenge and Workshop (AVEC 2011)*.

B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic, editors. 2011b. *Audio/Visual Emotion Challenge and Workshop (AVEC 2011)*.

F. Eyben M. Wollmer B. Schuller. 2009. Openear introducing the munich open-source emotion and affect recognition toolkit. In *ACII*.

N. Sebe, I. Cohen, T. Gevers, and T.S. Huang. 2006. Emotion recognition based on joint visual and audio cues. In *ICPR*.

D. Silva, T. Miyasato, and R. Nakatsu. 1997. Facial emotion recognition using multi-modal information. In *Proceedings of the International Conference on Information and Communications Security*.

S. Somasundaran, J. Wiebe, P. Hoffmann, and D. Litman. 2006. Manual annotation of opinion categories in meetings. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*.

P. Stone. 1968. *General Inquirer: Computer Approach to Content Analysis*. MIT Press.

C. Strapparava and R. Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007)*, Prague, Czech Republic.

M. Taboada, J. Brooke, M. Tofiloski, K. Voli, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(3).

R. Tato, R. Santos, R. Kompe, and J. M. Pardo. 2002. Emotional space improves emotion recognition. In *In Proc. ICSLP 2002*, pages 2029–2032.

Y.-I. Tian, T. Kanade, and J.F. Cohn. 2001. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97 –115, feb.

P. Turney. 2002. 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)*, pages 417–424, Philadelphia.

D. Ververidis and C. Kotropoulos. 2006. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, September.

J. Wagner, E. Andre, F. Lingenfelser, and Jonghwa Kim. 2011. Exploring fusion methods for multimodal emotion recognition with missing data. *Affective Computing, IEEE Transactions on*, 2(4):206 –218, oct.-dec.

X. Wan. 2009. 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.

J. Wiebe and E. Riloff. 2005. 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.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

M. Wiegand and D. Klakow. 2009. The role of knowledge-based features in polarity classification at sentence level. In *Proceedings of the International Conference of the Florida Artificial Intelligence Research Society*.

T. Wilson, J. Wiebe, and R. Hwa. 2004. Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of the American Association for Artificial Intelligence*.

M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll. 2010. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *IEEE Journal of Selected Topics in Signal Processing*, 4(5), October.

B. Yang and C. Cardie. 2012. Extracting opinion expressions with semi-markov conditional random fields. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea.

Z. Zhihong, M. Pantic G.I. Roisman, and T.S. Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *PAMI*, 31(1).