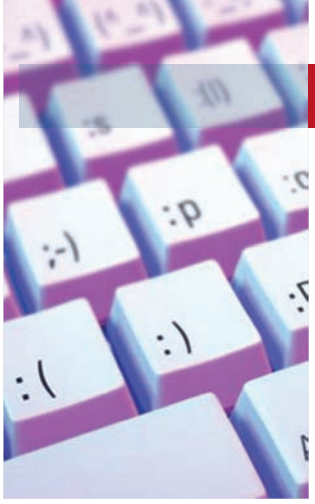

Editor: **Erik Cambria,** Nanyang Technological University, Singapore, cambria@ntu.edu.sg

Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages

Amir Zadeh, *Carnegie Mellon University*
Rowan Zellers, *University of Washington*
Eli Pincus, *University of Southern California*
Louis-Philippe Morency, *Carnegie Mellon University*

With the advent of mobile applications and social websites such as YouTube, Vine, and Vimeo, we have observed an increase in the number of online videos shared by people expressing their opinions, stories, and reviews. To give you a better idea how popular these websites are, more than 300 hours of video is uploaded to YouTube every minute. These videos address a large array of topics, such as movies, books, and products. This growth in multimedia sharing has seen increasing attention from many companies, researchers, and consumers interested in building better opinion-mining applications for summarization, question answering, and video retrieval. We highlight three challenges of studying sentiment in these online opinion videos.

The first challenge comes from the volatile and high-tempo nature of these opinion videos, wherein speakers often will switch between topics and opinions. This makes it challenging to identify and segment the different opinions expressed. For example, a speaker can express more than one opinion in the same spoken utterance, as in, "That was a great effect; there is a lot of cheap childish humor everyone can relate to, but I thought it was hilarious."

The second challenge comes with the range and subtlety of sentiment expressed in these opinion videos. We want approaches that can recognize the polarity of a video segment (for example, positive or negative) and also estimate the strength of the expressed sentiment.

The third challenge is a fundamental research question on how to use information more than text for sentiment analysis. In everyday communications, ideas and opinions are expressed through verbal content as well as visual and vocal behaviors, such as facial expressions, head gestures, and voice quality.

In this article, we introduce the Multimodal Opinion-Level Sentiment Intensity (MOSI) dataset, the video corpus with opinion-level sentiment intensity annotations that can be used for sentiment, subjectivity, and multimodal language studies. (For more information on text-based and multimodal sentiment analysis, see the "Background" sidebar.) We focus on psycholinguistic study of coverbal gestures.[1] Using a data-driven approach, we exploit prototypical interaction patterns between facial gestures and spoken words, and we introduce a new representation called *Multimodal Dictionary*. Finally, we evaluate our proposed Multimodal Dictionary on the challenging task of sentiment intensity prediction, using a speaker-independent paradigm (in which the model is tested on a new, unseen set of speakers to reduce the chance of bias introduced by speaker identification).

MOSI Dataset

In this section, we introduce our new MOSI dataset, the first such dataset to enable studies of multimodal sentiment intensity analysis. It can also be reliably used for detailed studies of language and gestures because of the rigorous annotation

1541-1672/16/$33.00 © 2016 IEEE
Published by the IEEE Computer Society

Background

Text-based sentiment analysis research has been an active and extremely successful field.[1] Among the notable efforts are works done in concept-level sentiment analysis,[2] automatic identification of opinion words and their sentiment polarity,[3] studies using *n*-grams and more complex language models,[4] works addressing sentiment compositionality by using polarity shifting rules or careful feature engineering,[5] and works that use deep learning approaches.[6] All these approaches primarily focus on the (spoken or written) text and ignore other communicative modalities.

Multimodal sentiment analysis has gained attention because of recent successes in multimodal analysis of human communications and affect.[7] Similar to our study are works that use support vector machines to classify sentiment polarity based on movie reviews,[8] that study multimodal sentiment analysis in Spanish videos,[9] that use convolutional neural networks and careful feature engineering for sentiment polarity classification,[10] and that use externally extracted word polarity data.[11] All the approaches in previous works use multimodal cues, including visual and acoustic cues. However, they have shortcomings with respect to core language and gestures studies, they present no analysis of sentiment intensity, and their approaches are speaker dependent.

References

1. B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *Proc. 42nd Ann. Conf. Assoc. Computational Linguistics*, 2004, article 271.
2. E. Cambria et al., "AffectiveSpace 2: Enabling Affective Intuition for Concept-Level Sentiment Analysis," *Proc. 29th AAAI Conf. Artificial Intelligence*, 2015, pp. 508–514.
3. M. Taboada et al., "Lexicon-Based Methods for Sentiment Analysis," *Computational Linguistics*, vol. 37, no. 2, 2011, pp. 267–307.
4. B. Yang, and C. Cardie, "Extracting Opinion Expressions with Semi-Markov Conditional Random Fields," *Proc. Jt. Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 1335–1345.
5. T. Nakagawa, K. Inui, and S. Kurohashi, "Dependency Tree-Based Sentiment Classification Using CRFs with Hidden Variables," *Proc. Ann. Conf. North Am. Chapter of the Assoc. Computational Linguistics Human Language Technologies*, 2010, pp. 786–794.
6. R. Socher et al., "Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank," *Proc. Conf. Empirical Methods in Natural Language Processing*, 2013, pp. 1631–1642.
7. E. Cambria, "Affective Computing and Sentiment Analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, 2016, pp. 102–107.
8. L.P. Morency, R. Mihalcea, and P. Doshi, "Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web," *Proc. 13th Int'l Conf. Multimodal Interfaces*, 2011, pp. 169–176.
9. V.P. Rosas, R. Mihalcea, and L.P. Morency, "Multimodal Sentiment Analysis of Spanish Online Videos," *IEEE Intelligent Systems*, vol. 28, no. 3, 2013, pp. 38–45.
10. S. Poria, E. Cambria, and A. Gelbukh, "Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-Level Multimodal Sentiment Analysis," *Proc. Conf. Empirical Methods in Natural Language Processing*, 2015, pp. 2539–2544.
11. S. Poria et al., "Fusing Audio, Visual and Textual Clues for Sentiment Analysis from Multimodal Content," *Neurocomputing*, vol. 174, 2016, pp. 50–59.

procedure. The dataset annotations contain the following:

- multimodal observations, including transcribed speech and visual gestures (an extensive set of automatically extracted text, audio, and visual features are also available for download with the dataset);
- opinion-level subjectivity segmentation;
- sentiment intensity annotations using unbiased crowdsourcing; and
- alignment between words, visual, and acoustic features.

The following subsections describe the dataset in more details.

Acquisition Methodology

We collected videos from YouTube with a focus on video blogs (vlogs)—popular monologue videos used by many YouTube users to express opinions about different subjects. The videos are recorded in diverse setups; some users have high-tech microphones and cameras, whereas others use less-professional recording devices. Users are in different distances from the camera with different lighting and background. The videos vary in length from 2 to 5 minutes. We selected a total of 93 videos from 89 distinct speakers, including 41 female and 48 male speakers. Most of the speakers were approximately between the ages of 20 and 30 years old. Although the speakers were from different ethnic backgrounds (for example, Caucasian, African American, Hispanic, and Asian), all speakers expressed themselves in English, and the videos originated from either the US or the UK. Figure 1 shows sample snapshots of video in the MOSI dataset.

We manually transcribed all the video clips to extract spoken words and the start time of each spoken utterance. Our transcription methodology had three stages. First, an expert transcriber manually transcribed all the videos, followed by a second transcriber reviewing and correcting all the transcriptions. Our transcription scheme contained details about pause fillers (such as "umm" and "uhh"), stresses, and speech pauses. In the third stage, the text was carefully aligned at word and phoneme levels with the videos using a forced alignment method called P2FA.[2] During the final stage, the results of the alignment were manually checked and, if necessary, corrected using PRAAT.[3]

Subjectivity Annotation

An important requirement of creating a dataset for sentiment analysis is to perform subjectivity segmentation to find opinionated segments of speech.

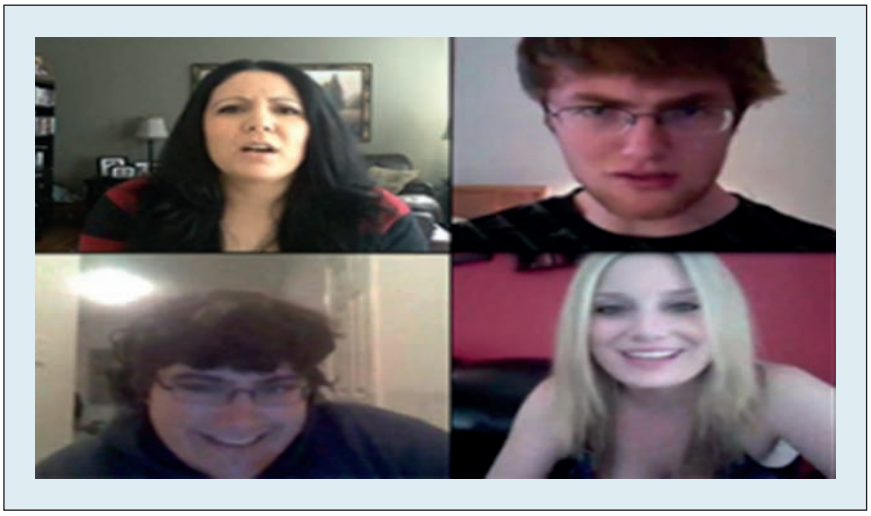

Figure 1. Example snapshots of videos from our new Multimodal Opinion-Level Sentiment Intensity (MOSI) dataset.

Table 1. MOSI dataset statistics.

Statistical measure	Value
Total no. segments	3,702
Total no. opinion segments	2,199
Total no. objective segments	1,503
Total no. videos	93
Total no. distinct speakers	89
Average no. opinion segments in video	23.2
Average length of opinion segments	4.2 seconds
Average word count per opinion segments	12
Total no. words in opinion segments	26,295
Total no. unique words in opinion segments	3,107
Total no. words in opinion segments appearing at least 10 times in the dataset	557

Following the work of Janyce Wiebe and colleagues,[4] *subjective sentences* are defined as expressions of a person's opinions, whereas *objective sentences* express facts and truth. Our annotation scheme expands their work to extract spoken opinion segments, defined to isolate distinct opinions and perform sentiment analysis on them. Therefore, subjective content comprises one or more opinion segments (hereafter, we will use *subjective segment*, *opinion segment*, and *opinion* interchangeably to refer to the same concept).

We define subjectivity as an attempt to express a private state, one that is distinguishable by carrying an opinion, belief, thought, feeling, emotion, goal, evaluation, or judgment. To more accurately annotate the boundaries of each opinion segment, we have defined the following rules. If the text contains an expression of a private state, the following segmentation rules apply (brackets are used to hold segments):

- Segment the subjective content on the basis of the number of private states revealed—for example, "[I love *The Shawshank Redemption*] [and I love *Transformers*]" results in two subjective segments.
- Segment if the utterance contains a modification of a private state while maintaining the subject—for example, "[Well, based on what I saw today, I feel like the movie industry is going crazy][or maybe it's just me being so hard on the poor actors.]"
- Segment if the subjective utterance ends with the start of an objective segment—for example, "[In my opinion, the movie was all about eating healthy food], you could see banners of different organic brands in several shots."

If there is subjective content and it extends beyond the boundary of the utterance while retaining the opinion, we merge the extension with the original utterance—for example, "[I don't like it! It's not a likable movie!]" The extension can be multiple sentences or part of a sentence.

Two trained annotators did the subjectivity annotation. The two annotations resulted in a Krippendorf's alpha of 0.68. The subjectivity annotation resulted in 2,199 subjective segments and 1,503 objective ones. We considered both subjective and objective segments for multimodal subjectivity studies, but for sentiment annotations, we focused on subjective segments. Table 1 gives detailed statistics of the dataset and opinion segments.

Crowdsourced Sentiment Intensity Annotation

Sentiment intensity is defined from strongly negative to strongly positive with a linear scale from –3 to +3. Online workers from Amazon Mechanical Turk performed the intensity annotations. Only master workers with an approval rate of higher than 95 percent were selected to participate. A total of 2,199 short video clips were created from the subjective opinion segments. For each video, the annotators had eight choices: strongly

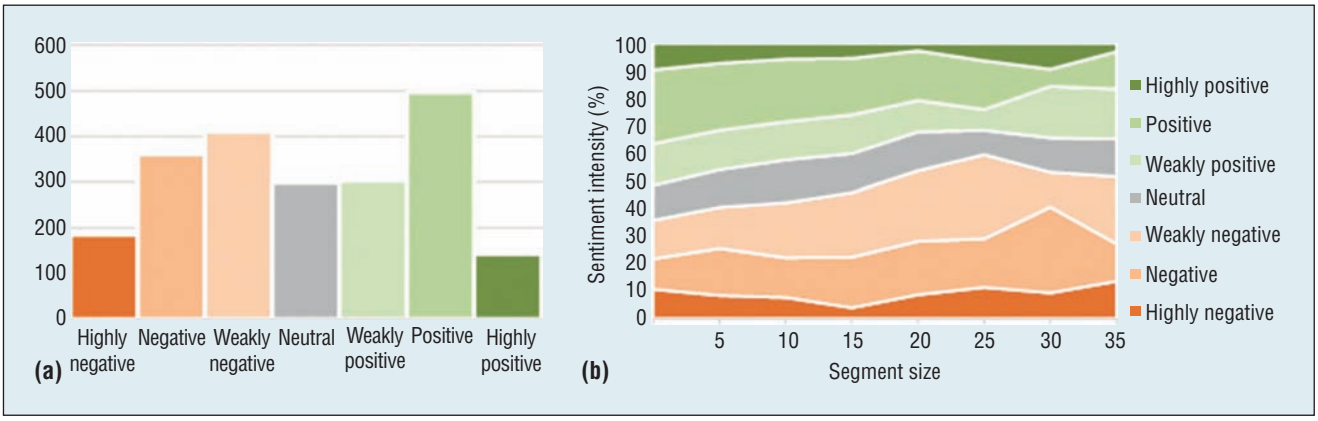

Figure 2. Histograms of sentiment distribution in MOSI dataset. (a) Distribution of sentiment over the entire dataset. (b) The percentage of each sentiment intensity per segment size (number of words in opinion segment).

positive (labeled as +3), positive (+2), weakly positive (+1), neutral (0), weakly negative (–1), negative (–2), strongly negative (–3), and uncertain.

We kept the instructions simple to reduce any training bias. The only tutorial was on how to use the online system (for example, how to submit the form). The task was phrased as follows: "How would you rate the sentiment expressed in this video segment? (Please note that you may or may not agree with what the speaker says. It is imperative that you only rate the sentiment stated by the speaker, not your opinion.)" Each video clip was annotated by five workers. The interannotator agreement between workers was 0.77 in terms of Krippendorf's alpha. The final sentiment intensity of each segment is the average of all five workers. Figure 2a shows the distribution of sentiment intensities for all opinion segments in the MOSI dataset; Figure 2b shows how the sentiment distribution changes as the size of the opinion (that is, the number of words in that opinion) increases. Although Richard Socher and colleagues reported that short (fewer than 10 words) text-only opinions have no significant sentiment and are mostly neutral,[5] short video segments have equal distribution along all scores, which shows that the presence of multimodal information, more than just text, makes it possible for human annotators to deduce sentiment for small opinion segments.

Manual Gesture Annotations

We provided a set of manually annotated gestures to study the relations between words and gestures. Because hands were not always visible in the YouTube videos, only facial gestures are annotated. We selected four gestures and expressions: smile, frown, head nod, and head shake. These are expressive of emotions and regularly happen in MOSI dataset. The annotations were done at the segment level. An expert coder manually annotated all 2,199 video segments, and a second coder annotated a subset of this dataset to compute the agreement between the coders. For all four gestures, the average coder agreement was 0.81.

Multimodal Analysis of Visual Gestures and Verbal Messages

The MOSI dataset enables detailed statistical study of language as a multimodal signal. We conducted a study to find a suitable multimodal representation for sentiment analysis. We wanted to understand the interaction patterns between spoken words and visual gestures. To study these interaction patterns, we studied the changes in the distribution of perceived sentiment intensity when a specific facial gesture is present or not. We performed this analysis at the opinion level, wherein we studied the multimodal interactions of the top 100 spoken words with all four facial gestures (smile, frown, head nod, and head shake).

Interaction Patterns

Figure 3 shows representative examples from our multimodal analysis, in which we identified four types of interaction patterns between spoken words and facial gestures: neutral, emphasizer, positive, and negative. Each subgraph is a histogram that represents the distribution of perceived sentiment intensities per opinion segment.

To help understand the average interaction of facial gestures with spoken words, the first row of Figure 3 shows how the sentiment intensities are distributed for all opinion segments (the top-left histogram of Figure 3 is repeated from Figure 2). It is not surprising to see that opinion segments with a smile or a head nod are perceived as more positive. The opposite effect is observed for frown and head shake gestures.

Neutral interaction pattern. To exemplify this pattern, we selected the most frequent word in our dataset: "the." "The" is considered sentimentally neutral in isolation. The second row of Figure 3 shows the interaction between the facial gestures and the spoken word

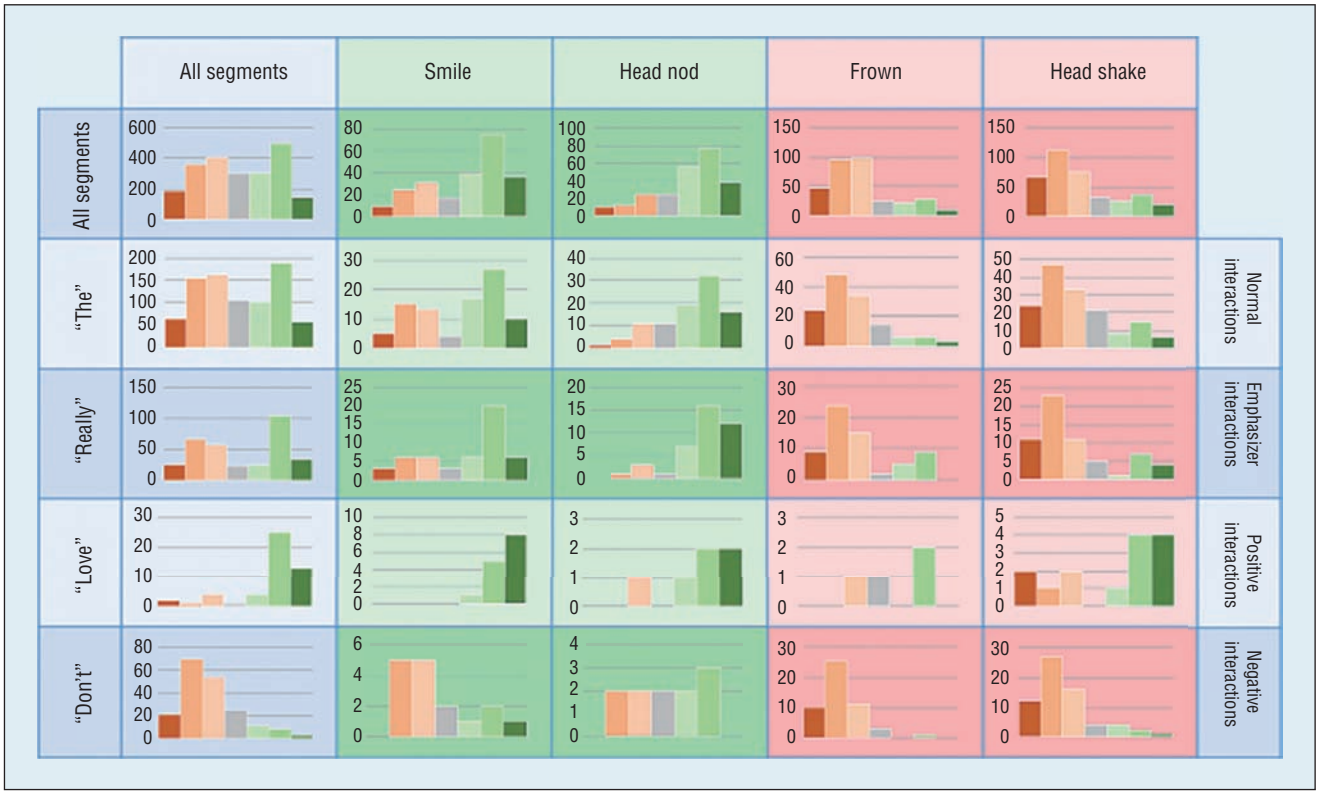

Figure 3. Sentiment intensity histograms for spoken words and visual gestures. In each histogram, the *y*-axis is the number of co-occurrence and the *x*-axis is the sentiment intensity as in Figure 2.

"the." We can observe that the pattern mostly follows the common interaction patterns in the first row.

Emphasizer interaction pattern. We observed a second interaction pattern in our multimodal analysis. To exemplify this pattern, we showed in the third row of Figure 3 how facial gestures interact with the word "really." When accompanied by a smile or head nod, the distribution tends to shift to positive sentiment intensity, with less negative or neutral intensities. The opposite effect happens when "really" is accompanied by a frown or head shake—then, the distribution is biased toward negative sentiment. In other words, this interaction pattern tends to shift the sentiment toward the extremes. We define this type of interaction pattern as an *emphasizer*.

Negative and positive interaction patterns. The third and fourth type of interaction seem to appear when studying sentimentally polarized words, because their sentiment distributions are not affected in the same way as the neutral and emphasizer. For example, the positively polarized word "love" is shown in the fourth row of Figure 3 (we merged the verbs "love" and "loved" in these histograms for simplification). The sentiment distributions do not significantly change polarity when accompanied by a frown or head shake. An opposite trend happens when we study a negatively polarized word such as "don't," shown in the last row of Figure 3 (we merged all instances of "don't," "doesn't," and "didn't" in these histograms). We observe limited changes in the sentiment distributions for the smile and head nod.

Multimodal Dictionary

On the basis of these interaction patterns between words and gestures, we present a simple representation model that jointly accounts for words and gestures in each opinion segment. We define $W = \{w_1, w_2, w_3, \ldots, w_K\}$ as the set of words in our dataset and K as the dictionary size. We observed in our experiments that information about gestures being present or not present is useful; thus, we defined $G = \{$*smile*, *frown*, *head nod*, *head shake*, *~smile*, *~frown*, *~head nod*, *~head shake*$\}$ (the approximation symbol indicates no evidence of that gesture). We then defined the Multimodal Dictionary to be the Cartesian product of sets of words W and gestures G as follows:

$$M = \{(w, g) \mid w \in W, g \in G\}.$$

The Multimodal Dictionary creates a simple joint space of words and gestures. Each element in this multimodal representation is a binary variable similar to the bag-of-words representation for text and captures if a word

and gesture have co-occurred. Using this method yields better results in sentiment intensity analysis compared with common fusion methods.

Experimental Results

All the experiments described in this section were done in a speaker-independent framework. We trained prediction models using nu-SVR[6] and tested them using a fivefold cross-validation methodology. The automatic validation of the hyperparameters was performed with fourfold cross-validation on the training sets. We calculated the regressors' performance based on mean absolute error (MAE) and correlation. We trained the following models:

- *Random*. We included in our experiments a simple baseline model that always predicts a random sentiment intensity between [3, –3]. This baseline gives an overall idea about how random models will work.
- *Verbal*. We trained this model using only verbal features from MOSI. We created a simple bag-of-words feature set from monograms and bigrams created from words in speech segments, including speech pauses and pause fillers. All the features with fewer than 10 instances in the dataset were removed from the bag-of-words set, given their infrequency.
- *Visual*. We trained this model using facial gestures, as described earlier. We assigned a binary feature for each of the four facial gestures: smile, frown, head nod, and head shake.
- *Verbal+Visual*. We trained this model on verbal and visual data combined. The verbal and visual features were simply concatenated for each opinion segment.
- *Multimodal Dictionary*. We trained this model on Multimodal Dictionary representation. Each element in the Multimodal Dictionary is treated as a random variable and denotes joint representation between words and gestures.
- *Human Performance*. Humans are asked to predict the sentiment score of each opinion segment. This will be both a baseline for how well humans can predict sentiment intensity and a future target for machine learning methods.

Table 2. Mean absolute error and correlation for each of the trained baseline models.

Model	Mean absolute error	Correlation
Random	1.88	0.00
Verbal	1.18	0.46
Visual	1.24	0.36
Verbal+Visual	1.14	0.49
Multimodal Dictionary	**1.10**	**0.53**
Human Performance	0.61	0.83

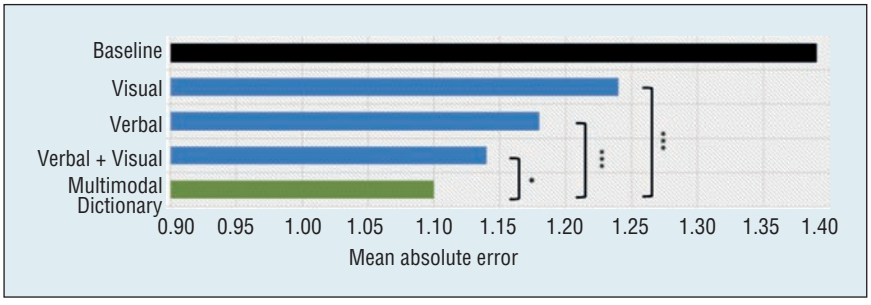

Figure 4. Statistical comparison between Multimodal Dictionary and other trained models. One star shows $p < 0.01$, and three stars show $p < 0.0001$.

Table 2 summarizes our experimental results. We performed a significance test using pairwise *T*-test between the models (see Figure 4; stars indicate the *p*-value range). Our first observation from these results is that the Verbal+Visual model outperforms both the Verbal model and Visual model individually.

A second observation is that the Multimodal Dictionary model outperforms the Verbal+Visual model. The difference is statistically significant ($p < 0.01$). This is well-aligned with multimodal study mentioned earlier and presented in Figure 3, wherein spoken words and facial gestures were shown to have multiple interaction patterns. Our Multimodal Dictionary is designed to explicitly model these interactions. This new representation results in better performance for sentiment intensity prediction.

In this article, we introduced the Multimodal Dictionary to better understand the interaction between facial gestures and spoken words when expressing sentiment. This new computational representation improved prediction performance in speaker-independent multimodal sentiment intensity analysis. The findings we present here open the door to new research directions for studying human communication dynamics. One promising future direction is to analyze the vocal behaviors (such as vocal emphasis or prosodic cues) in the context of multimodal sentiment expressions. This analysis should also be augmented to take into account the temporal contingency between these vocal, visual, and verbal behaviors. These research directions will help us better under-

stand the dynamics of human communications central to many applications such as healthcare and education. ■

Acknowledgments

This material is based on work supported in part by the National Science Foundation under grant no. IIS-1118018 and Yahoo Research. The content does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

References

1. D. McNeill, *Language and Gesture*, vol. 2, Cambridge Univ. Press, 2000.
2. J. Yuan, and M. Liberman, "Speaker Identification on the SCOTUS Corpus," *J. Acoustical Soc. Am.*, vol. 123, no. 5, 2008, pp. 3878.
3. P. Boersma, "Praat, A System for Doing Phonetics by Computer," *GLOT Int'l*, vol. 5, no. 9/10, 2002, pp. 341–345.
4. J. Wiebe, T. Wilson, and C. Cardie, "Annotating Expressions of Opinions and Emotions in Language," *Language Resources and Evaluation*, vol. 39, nos. 2–3, 2013, pp. 165–210.
5. R. Socher et al., "Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank," *Proc. Conf. Empirical Methods in Natural Language Processing*, 2013, pp. 1631–1642.
6. A.J. Smola, and B. Schölkopf, "A Tutorial on Support Vector Regression," *Statistics and Computing*, vol. 14, no. 3, 2004, pp. 199–222.

Amir Zadeh is a PhD student in deep learning at Carnegie Mellon University and a Yahoo InMind Fellow. Contact him at abagherz@cs.cmu.edu.

Rowan Zellers is a PhD student in computer science at the University of Washington. Contact him at rowanz@uw.edu.

Eli Pincus is a PhD student in spoken dialogue at the University of Southern California and a research assistant in the Natural Dialogue Group at USC's Institute for Creative Technologies. Contact him at pincus@ict.usc.edu.

Louis-Philippe Morency is an assistant professor in the Language Technology Institute at Carnegie Mellon University, where he leads the Multimodal Communication and Machine Learning Laboratory. Contact him at morency@cs.cmu.edu.