Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web

Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
Los Angeles, CA 90094
morency@ict.usc.edu

Rada Mihalcea
University of North Texas
Denton, TX 76203
rada@cs.unt.edu

Payal Doshi
University of Southern California
Los Angeles, CA 90094
doshi@usc.edu

ABSTRACT

With more than 10,000 new videos posted online every day on social websites such as YouTube and Facebook, the internet is becoming an almost infinite source of information. One crucial challenge for the coming decade is to be able to harvest relevant information from this constant flow of multimodal data. This paper addresses the task of multimodal sentiment analysis, and conducts proof-of-concept experiments that demonstrate that a joint model that integrates visual, audio, and textual features can be effectively used to identify sentiment in Web videos. This paper makes three important contributions. First, it addresses for the first time the task of tri-modal sentiment analysis, and shows that it is a feasible task that can benefit from the joint exploitation of visual, audio and textual modalities. Second, it identifies a subset of audio-visual features relevant to sentiment analysis and present guidelines on how to integrate these features. Finally, it introduces a new dataset consisting of real online data, which will be useful for future research in this area.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Discourse*

General Terms

Algorithms, Experimentation

Keywords

Multimodal signal processing, Subjectivity and sentiment analysis, Audio-visual integration, YouTube videos

ICMI'11, November 14–18, 2011, Alicante, Spain.

1. INTRODUCTION

Subjectivity and sentiment analysis focuses on the automatic identification of private states, such as opinions, emotions, sentiments, evaluations, beliefs, and speculations in natural language. While subjectivity classification labels data as either subjective or objective, sentiment classification adds an additional level of granularity, by further classifying subjective data as either positive, negative or neutral.

Much of the work to date on subjectivity and sentiment analysis has focused on textual data, and a number of resources have been created including lexicons [28] or large annotated datasets [18, 29]. 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, Facebook), audio (e.g., podcasts), the ability to address the identification of opinions and sentiment for diverse modalities is becoming increasingly important. With only one exception [20], we are not aware of any previous work that attempted to combine multiple modalities for the purpose of opinion and sentiment analysis. Moreover, although there is a significant amount of previous work on multi-modal emotion analysis, that work has not addressed specifically the polarity (or sentiment) of data, and has generally focused on visual and audio cues, and mainly ignored the knowledge that can be gathered from textual analysis.

In this paper, we address the task of multimodal sentiment analysis, and conduct proof-of-concept experiments that demonstrate that a joint model that integrates visual, audio, and textual features can be effectively used to identify sentiment in web data. Specifically, this paper makes three important contributions. First, we address for the first time the task of tri-modal sentiment analysis by integrating three different modalities: visual, audio and linguistic features, which are jointly used to determine the polarity of an input stream. This is unlike most of the work done on multimodal emotion analysis, which often addressed only one or two modalities at a time (e.g., visual and audio cues). Second, we present a qualitative and statistical analysis that identifies five multimodal features that are found helpful to differentiate between negative, neutral, and positive sentiments: polarized words, smile, gaze, pauses, and voice pitch. Finally, in our experiments we target and use real online

data, which poses additional challenges with respect to the artificial datasets that have been typically used in the past in multimodal research. We introduce a new dataset consisting of video opinions, collected from the YouTube web site, which we analyse and annotate for sentiment. The results of our initial experiments show that the joint use of multiple modalities can improve significantly over classifiers that use only one modality at a time, thus demonstrating the potential of multimodal sentiment analysis.

The paper is organized as follows. We first review related work on sentiment and emotion analysis, followed by a description of the problem of multimodal sentiment analysis in Section 3. Section 4 introduces the new dataset that we use in the experiments, including a description of the data acquisition, transcription, and annotation. Section 5 describes our qualitative and statistical analysis of several audio-visual features relevant to sentiment analysis. Section 6 presents our multimodal sentiment classifier, our multimodal features, followed by experimental methodology and a discussion of the results.

2. PREVIOUS WORK

Our paper follows a novel approach of combining audio-visual and text features from videos for sentiment analysis. It evolved from thoughtful reading of many different methods followed previously for audio-visual emotion recognition or text based sentiment analysis. Each of these methods involves a different approach, dealing with either the extraction and processing of data, or with the machine learning algorithms used for the classification.

To date, a large number of text processing applications have already used techniques for automatic sentiment and subjectivity analysis, including automatic expressive text-to-speech synthesis [1], tracking sentiment timelines in online forums and news [3], and mining opinions from product reviews [14]. In many natural language processing tasks, subjectivity and sentiment classification has been used as a first phase filtering to generate more viable data. Research that benefited from this additional layering ranges from question answering [31], to conversation summarization [8] and text semantic analysis [27].

The techniques developed so far for sentiment analysis focus primarily on the processing of text, and consist of either rule-based classifiers that make use of sentiment lexicons, or data-driven methods that assume the availability of a large dataset annotated for polarity. For instance, one of the most frequently used lexicons is the subjectivity and sentiment lexicon provided with the OpinionFinder distribution [28], which contains 6,856 unique entries that are also associated with a polarity label, indicating whether the corresponding word or phrase is positive, negative, or neutral. Another lexicon that has been often used in polarity analysis is the General Inquirer [24], which is a dictionary of about 10,000 words grouped into about 180 categories that have been widely used for content analysis. Two of the largest categories in the General Inquirer are the valence classes, which form a lexicon of 1,915 positive words and 2,291 negative words. SentiWordNet [10] is a resource for opinion mining built on top of WordNet, which assigns each synset in WordNet with a score triplet (positive, negative, and objective), indicating the strength of each of these three properties for the words in the synset.

Another major line of work in sentiment and subjectivity analysis consists of data-driven methods based on annotated corpora. One of the most widely used datasets is the MPQA corpus [29], which is a collection of 535 English-language news articles from a variety of news sources manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.). The corpus was originally annotated at clause and phrase level, but sentence-level annotations associated with the dataset can also be derived via simple heuristics [28]. Another manually annotated corpus is the collection of newspaper headlines created and used during the Semeval task on "Affective Text" [25], which consists of 1000 test headlines and 200 development headlines, each of them annotated with the six Eckman emotions (anger, disgust, fear, joy, sadness, surprise) and their polarity orientation (positive, negative). Two other data sets, both of them covering the domain of movie reviews, are a polarity data set consisting of 1,000 positive and 1,000 negative reviews, and a subjectivity data set consisting of 5,000 subjective and 5,000 objective sentences. Both data sets have been introduced in [18], and have been used to train opinion mining classifiers.

Building upon these or other related resources, there is a growing body of work concerned with the automatic identification of sentiment in text, which often times addresses online text, such as reviews [18],news articles [3], or blogs [12]. While difficult problems such as cross-domain [5] or cross-language [16]portability have been addressed, not much has been done in terms of extending the applicability of sentiment analysis to other modalities, such as speech, gesture, or facial expressions. The only exception that we are aware of is the research reported in [20], where speech and text have been analyzed jointly for the purpose of opinion identification. This previous work, however, did not address other modalities such as visual cues, and did not address the problem of sentiment analysis.

Also related to our work is the research done on multimodal emotion analysis [7, 23, 32]. For instance, a novel algorithm is defined in [30], based on a combination of audio-visual features for emotion recognition. The features used by these novel algorithms are usually basic and low level like tracking points for collecting visual data. An engineering approach is then applied to this large set of data points, in order to extract the ones that would be useful for the actual analysis. To our knowledge, there is no previous work analyzing sentiments using all three modalities: textual, audio and video.

Further, most sentiment analysis datasets are created by the researchers in the scientific environment with accurate precision recording and reduced noises [6, 4, 11]. In real world data, however, there are many additional difficulties to overcome. The real world data is the one recorded by people on their own with household instruments like an inbuilt microphone system and an inbuilt web cam of a laptop. Such data would typically have a much greater amount of noise, and the algorithms that work well on scientifically recorded data might prove inefficient on the real world data.

3. MULTIMODAL SENTIMENT IN WEB VIDEOS

Sentiment analysis has recently become a new trend in social media, where it helps users (whether they are brands or consumers) to understand the opinions being expressed

about events, products, people, locations, etc. Most of these opinions are given voluntarily and hence are considered to be honest feedback. With the advancement of technology and its increased use amongst masses, in addition to the large body of opinions expressed in textual format, there is a growing number of opinions that are available in video format. Thus, in addition to sentiment analysis for textual data, we also need means to analyze multimodal data, and thereby understand the large number of recorded videos where opinions are being expressed.

Consumers tend to record videos using their web cams or similar devices, to express their opinion on various topics. The main goal is to let other people know about their personal experiences regarding a particular event, product or entity, which they would like to share with others. These videos might include a discussion of the user opinions on a certain topic; a comparison of various brands or products; a discussion of the strengths and weaknesses of a specific product; etc. All these categories are equally helpful to understand the consumer needs and opinions.

These opinion videos are often openly available to all who need to know about a certain topic, through various open Web sources such as YouTube or Facebook. A person who needs to buy a product would first look at the reviews by people already using it. Similarly, someone who tries to determine the movie to watch on a Saturday night will first check the reviews available for that movie. Videos are known to have maximum impact on a person's views, and hence such video feedbacks or opinions have the potential to make a strong impact over the consumer market and are important for understanding the consumer expectations and needs.

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

The video data can be a good source for sentiment analysis or opinion mining but it also comes with many challenges that need to be addressed before using the freely available video opinions. The expressiveness of emotions varies from person to person. Some people express themselves more vocally while others more visually. A person with more vocal modulation will have most of the data that would be necessary for opinion mining stored in audio responses. On the other side, a person whose expressing herself more visually would have most of the data stored in facial expressions and non-verbal cues that need to be recognized. A generic model should be able to adapt itself based on each user and give a consistent result.

Thus, the main challenges consist of the noise that is often present in such online, multimodal data, as well as the differences in person-to-person communication patterns. Both these conditions result in difficulties that need to be addressed in order to effectively extract useful data from these sources.

Figure 1: Selected snapshots from our new video dataset.

4. YOUTUBE DATASET

As discussed in the previous section, the automatic analysis of sentiment and subjectivity from real-world videos is a challenging research problem. To properly address this new research direction, we created a dataset[1] from online social videos that encompasses the different facets of sentiment analysis:

- **Diversity:** Diversity comes in multiple dimensions when analyzing real-world interactions. People express sentiments in multiple ways, and some people will be more subtle than others. Also, the topics addressed in online social videos are extremely diverse (e.g., religion views, politic opinions, product reviews). Our dataset contains videos of people from different age and gender groups, expressing opinions on diverse topics.
- **Multimodal:** While an image is worth a thousand words, selecting the appropriate word (e.g., gorgeous, ridiculous) can get you a long way in expressing a sentiment. Through a mixture of facial expressions, body postures, intonations and choice of words, people are extremely efficient at expressing different sentiments. Our dataset contains multimodal videos where one person is speaking directly at the camera, expressing their opinion and/or stating facts.
- **Ambient noises:** A robust computational model of sentiment analysis needs to be able to handle the real-world variability and noises present in most video recordings. While the previous research on audio-visual emotion analysis used videos recorded in laboratory settings[4, 6, 11], our dataset contains videos recorded by users in their home, office or outdoor, using different web cameras and microphones.

Social media Web sites such as YouTube are the perfect place to acquire such an interesting dataset. In fact, more than 10,000 videos are added to YouTube every day. People from all around the world post videos online and these videos are freely available (given proper acknowledgment of publishing licenses). Also, social media Web sites contain the diversity, multi-modality and ambient noises characterizing real-world sentiment analysis.

The following section describes how we acquired our dataset and Sections 4.3 and 4.2 present our approaches to transcribe and annotate the videos. The details about the automatically extracted multimodal features are described later in the Experiment section (see Section 6.2).

[1]Please visit the following website for the video links, transcriptions and sentiment annotations: http://projects.ict.usc.edu/youtube/

4.1 Acquisition

We collected 47 videos from the social media web site YouTube. As mentioned earlier, an important characteristic of our dataset is its generalized nature. The dataset is created in such a way that it is not based on one particular topic. The videos were found using the following keywords: opinion, review, product review, best perfume, toothpaste, war, job, business, cosmetics review, camera review, baby product review, I hate, I like.

The final video set has 20 female and 27 male speakers randomly selected from youtube.com, with their age ranging approximately from 14-60 years. Although from different ethnic backgrounds (e.g., Caucasian, African-American, Hispanic, Asian), all speakers expressed themselves in English. The videos are converted to .mp4 format with a standard size of 360x480. The length of the videos varies from 2-5 minutes.

All videos are pre-processed to address the following issues: introductory titles and multiple topics. Many videos on YouTube contain an introductory sequence where a title is shown, sometime accompanied with a visual animation. As a simple way to address this issue, the first 30 seconds of each video is removed. In the future, this step could be optimized by automatically performing optical character recognition (OCR) on the videos [19].

The second issue is related to multiple topics. Videos posted on social networks can address more than one topic. For example, a person can start by talking about the new movie she recently saw and then switch to a new (or related) topic such as food served in movie theaters. To simply address this issue, all video sequences are normalized to be 30 seconds in length. We keep as future work to automatically segment topics based on transcriptions [2] or directly based on the audio-visual signals.

4.2 Transcriptions

All video clips are manually transcribed to extract spoken words as well as the start time of each spoken utterance. The Transcriber software was used to perform this task. The transcription is done using only the audio track of each video clip. In other words, the transcriptions are done without the visual information. Although not used in this paper, the word elongations and filler pauses are also transcribed. Each video contains 3-11 utterances with most videos having 5-6 utterances in the extracted 30 seconds. The utterance segmentation was based on long pauses which could easily be detected using tools such as Praat and OpenEAR [22].

Multimodal sentiment analysis using manual transcription is a precedent step to fully automatic sentiment classification. This paper is a proof-of-concept that tri-modal sentiment classification can be performed. Automatic speech recognition can also be used as input to our approach. In the recent years, many technologies have emerged to automatically transcribe voicemails (e.g., Google Voice) and videos (e.g., Adobe Translator). In fact, not all words need to be accurate since we are using a dictionary of polarized words and valence shifters, and thus only the words in the dictionary need to be properly recognized.

4.3 Sentiment Annotations

An important step in creating such a dataset is sentiment label annotation. Since our goal is to automatically find the sentiment expressed in the video clip, we decided to perform our annotation task at video sequence level. For each video, we want to assign one of these three labels: negative, neutral or positive. All 47 video clips are annotated by three annotators who were shown videos in three different random sequencing orders, so as to reduce the compound effect. It is important to note that we are not annotating the sentiment felt by the person watching the video. The annotation task is to associate a sentiment label that best summarizes the opinion expressed in the YouTube video.

To perform this annotation task, we built a web interfaces that shows one video clip at the time and the three label options. The annotators can replay the video as often as they want. Once they are ready, they select the sentiment label and go to the next video clip.

The pair-wise coder agreements are 89.4%, 82.9% and 63.8% respectively, with an average coder agreement of 78.7%. None of the videos had a complete disagreement where all three coders disagreed. Always two out of three annotators agreed. Based on this observation, we define our final ground truth labels using majority voting. Out of the 47 videos clips, 13 were labeled as positive, 22 as neutral, and 12 as negative.

5. SENTIMENT ANALYSIS EXPERIMENTS

The study of human verbal and nonverbal behaviors when interacting with social medias such as Skype and Youtube is an ongoing research topic. Much needs to be analyzed to completely understand the influence of these new technologies on human multimodal interactions. The goal of this paper is not only to create technology to automatically classify sentiments (as described in Section 6) but also to gain a better understanding of how nonverbal cues accompany positive and negative sentiments. This section presents a first step in this direction where we analyze the nonverbal behaviors co-occurring with negative, neutral and positive videos.

5.1 Qualitative Analysis

We started our analysis by observing qualitatively the videos of our corpus, searching for nonverbal cues that could be correlated with sentiments. In this first stage, we look for audio-visual cues that vary among people present in our 47 videos. We focus on facial motion and the basic prosodic cues (voice intensity and pitch).

Please note that a more engineering approach could have been used to perform this pre-filtering step. In fact, many researchers from the signal processing community approach this problem by simply trying all possible features under all possible representations [30]. While this is a valuable approach, these models often end up really large (containing multiple features) and are harder to interpret. In contrast, we decided to start with a more behavioral approach where we first studied the videos directly, searching for relevant nonverbal cues. Applying the engineering approach to our YouTube dataset will be a great future research direction but it is not necessary at this point.

From this qualitative pre-study, four audio-visual cues were identified as potential features: smile, gaze, pauses and voice pitch. In all four cases, we observed large variations which could possibly be correlated with sentiment expression. The details related to tri-modal feature extraction are described in Section 6.2. The following present the results of our statistical analysis.

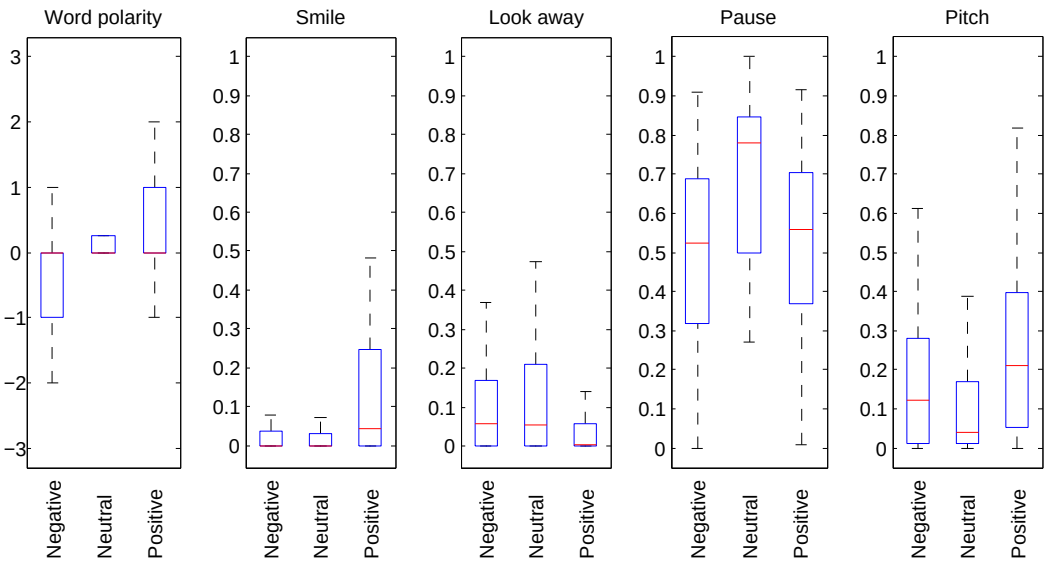

Figure 2: Average values of multimodal features when clustered per sentiment label. In all five graphs, the red line represents the median, the top blue line represents the 75th percentile and the bottom blue line represents the 25th percentile. Word polarity is a great way to differentiate sentiment but many utterances do not contain polarized words as shown by all three medians equal to zero. The visual features (smile and look away) are good ways to differentiate positive utterances from neutral or negative utterances. Audio features (pauses and pitch) are great ways to differentiate neutral utterances from positive or negative utterances.

5.2 Statistical Analysis

To confirm a correlation between the observed nonverbal cues and the sentiment expressed by the person, we perform a statistical analysis based on percentile ranking. This analysis will give us a better insight on which nonverbal and verbal cue is relevant to identify a specific sentiment (negative, neutral and positive). We will later use these results to learn a computational model and automatically classify sentiments in web videos.

Using the audio-visual features identified in the previous section (see details in Section 6.2), we perform percentile ranking per sentiment labels. In other words, we look at variation of each multimodal features for each sentiment label. The sentiment labels come from the annotations described in Section 4.3.

Figure 4.1 shows the results of the percentile ranking. In all five graphs, the red line represents the median, the top blue line represents the 75th percentile and the bottom blue line represents the 25th percentile. From these results, many interesting observations can be made:

- **Polarized words:** As shown in previous work on text-based sentiment analysis[15, 26], using a dictionary of positively or negatively polarized words, sentiment polarity can be effectively differentiated. One of the main issues with using only textual features is that most utterances do not contain polarized words. This is shown in Figure 4.1, where all three medians of Polarized Words are equal to zero.

- **Smile:** The fact that people smile when expressing positive sentiment does not come as a surprise. Smile has been shown to be correlated with happiness [9]. Smile is a good feature to differentiate positive utterances from neutral or negative utterances.

- **Look away:** People gaze patterns seem correlated with sentiment. We can see that people look away from the camera more often when expressing neutral or negative utterances. Another way to interpret this result is to see that people try to create mutual-gaze more often when expressing positive utterances. In this case, mutual-gaze is approximated by having the speaker looking at the camera. It was shown that people have more mutual gaze when they are creating rapport and engagement [13], which can be seen as more positive.

- **Pauses:** Our analysis show that people seem to pause less often when expressing polarized (negative or positive) utterances. Often these polarized utterances are spoken with more energy and faster pace. So we see more pauses with neutral utterances.

- **Pitch:** A similar pattern emerged with the voice pitch where polarized utterances are spoken with a wider range of variation in the pitch level. The neutral utterance are spoken at a more monotone level.

In summary, polarized words are great to differentiate sentiment although many utterances do not contain polarized words. The visual features are good ways to differentiate positive utterances from neutral or negative utterances. Audio features are great ways to differentiate neutral utterances from positive or negative utterances. By combining all three modalities we expect a better classification performance.

6. SENTIMENT CLASSIFICATION EXPERIMENTS

Our final goal is to be able to automatically classify an audio-visual clip into one of these three sentiment labels: positive, negative, or neutral. The experiment described in this section is based on the expertise and knowledge learned from the qualitative and statistical analysis described in the previous section. We know that textual, audio and visual features can be used to differentiate sentiment in spoken data. The remaining question is whether these modalities are complementary, and if so, can they help each other in the classification?

To test this hypothesis, we train a computational model of spoken sentiments and test it using our Youtube dataset. The following three sections describe our computational model, the feature extraction and the experimental methodology, while Section 6.4 presents and discusses our results.

6.1 Computational Model

Human communication is a dynamic process and explicitly modeling this dynamic has been shown to improve tasks such as gestures recognition [17] and speech recognition [21]. One probabilistic model which was shown to perform well in these situations is the HMM classifier. This model learns the hidden structure in the input signal by jointly learning the observation model (i.e., the relationship between the input features and the hidden variables) and the dynamic (or motion) model.

In our case, the dynamic is most prominent at the utterance level. We are thus learning HMMs that take as input tri-modal features summarizing each utterance. In other words, we model the YouTube video clip using a Markov chain where each element of the chain represents one spoken utterance (see Section 4.2 for details about the utterance segmentation). For each utterance, we compute tri-modal features (described in the following section) summarizing the audio-visual cues happening during this utterance. The observation and dynamic models of our tri-modal HMM learn the relative importance of each audio-visual cues as well as the dynamic between utterances. For example, people often insert one or two negative points during a positive opinion to show that they considered both the positive and the negative aspects of issue being discussed. We use a mixture of Gaussian as our observation model and two hyper parameters are automatically validated: the number of Gaussian mixtures and the number of hidden states.

6.2 Automatic Feature Extraction

Given an video clip with transcribed words (e.g., captions), we want to automatically extract multimodal features for sentiment analysis. Based on the literature and our own qualitative analysis, we define five important tri-modal features: polarized words, smile, look-away, pauses, and voice pitch. Following previous work on audio-visual emotion recognition [30, 7], we extract features at the utterance level so that we can represent local changes. Our multimodal sentiment classifier described in Section 6 will learn the hidden dynamic between utterances and output sentiment labels at the video level. The following three subsections describe how we automatically extract these multimodal features.

6.2.1 Text features

We generate textual features by automatically identifying linguistic cues of sentiment present in the text of the utterance. First, using two lexicons of words labeled as positive or negative, the presence of sentiment words in the text is identified, and a polarity score is calculated. The lexicons are compiled from a distribution of the MPQA dataset (Wiebe 2005), and consists of words loaded with positive (e.g., "good") or negative (e.g., "bad") sentiment. We assign each word in these lexicons with a predefined polarity value; for instance, a positive word could be assigned with a value of 1, and a negative word could be assigned with a value of -1. Other polarity values can also be used, and can lead to different accuracy figures in the automatic annotation of sentiment.

Next, we use a lexicon of valence shifters, which can change the polarity of a word (and correspondingly its polarity value) if a valence shifter is found within a certain distance. For instance, if the valence shifter "not" is found within two words from the positive word "good," the polarity of "good" is shifted from positive to negative, and its polarity value is changed from 1 to -1.

Given these lexicons, the polarity score of a text is calculated as the sum of the polarity values of the lexicon words found in the text, while also accounting for the valence shifters found in the text within close proximity of the lexicon words. The current system implementation defines "proximity" as words found within a distance of at most two words, but other window sizes can be used.

6.2.2 Visual features

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

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

- **Smile duration:** Given the start and end time of an utterance, how many frames are identified as "smile." In our experiments, we use three different variants of this feature with different thresholds: 50 and 75.
- **Look-away duration:** Given the start time and end time of the utterance, in how many frames is the speaker looking at the camera. The horizontal and vertical angular thresholds were experimentally set to 10 degrees.

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

6.2.3 Audio features

The audio features are automatically extracted from the audio track of each video clip. The audio features are extracted at the same framerate as the video features (30Hz) with a sliding window of 50ms. We used the open source software OpenEAR [22] to automatically compute the pitch and voice intensity. Speaker normalization is performed using z-standardization. The voice intensity was simply thresholded to identify samples with and without speech. The same threshold was used for all the experiments and all the speakers.

For each utterance in our dataset, we define two summary features:

- **Pause duration:** Given the start and end time of the utterance, how many audio samples are identified as silence. This audio feature is then normalized by the number of audio samples in this utterance. This feature can be interpreted as the percentage of the time where the speaker was silent.

- **Pitch:** Compute the standard deviation of the pitch level for the spoken utterance. This measure represents the variation of voice intonation during the same utterance.

These tri-modal features (polarized words, smile, gaze, pause and pitch) were used for both the statistical analysis described in the next section as well as the sentiment classification experiment described in Section 6.

6.3 Methodology

We perform leave-one-out testing where one video clip is kept for testing and all remaining 46 clips are used for training and validation. This process is repeated 47 times. All models described before are trained and tested using the same procedure. The optimal hyper-parameters (number of hidden states and number of mixtures) were automatically validated using the training error. The number of hidden states was validated with values 2, 3, 4, 5, 6 and 7 respectively. Four different number of Gaussian mixtures were tested for the HMM model: 1, 2, 3 and 4.

The performance is measured by using the F-measure, which is the weighted harmonic mean of precision and recall. Precision is the probability that predicted backchannels correspond to actual listener behavior. Recall is the probability that a backchannel produced by a listener in the test set was correctly predicted by the model. We use the same weight for both precision and recall, resulted in the so-called F1 measure. The same evaluation metric is applied to all the models. The training of all the HMMs models is done using the BNT Matlab toolbox from Kevin Murphy.[2]

6.4 Results

As stated earlier, the main goal of this experiment is to test if the multimodal features identified in Section 5 are able to work together to improve sentiment classification or if they are simply redundant features.

Table 1 shows the classification performances of four different models: text-only, visual-only, audio-only and tri-modal integration. These results show a significant improvement when all three sources of information are integrated, and, importantly, these improvements are observed for both precision and recall. The tri-modal HMM is able to learn the hidden interaction between all three modalities and take advantage of their respective discriminative power.

[2]http://code.google.com/p/bnt/

	F1	Precision	Recall
Text only HMM	0.430	0.431	0.430
Visual only HMM	0.439	0.449	0.430
Audio only HMM	0.419	0.408	0.429
Tri-modal HMM	**0.553**	**0.543**	**0.564**

Table 1: Automatic sentiment classification performances of four different models: text-only, visual-only, audio-only and tri-modal integration.

7. DISCUSSION AND CONCLUSIONS

In this paper, we addressed the task of multimodal sentiment analysis, and explored the joint use of multiple modalities for the purpose of classifying the polarity of opinions in online videos. Through experiments performed on a newly introduced dataset, consisting of videos where people express their opinion about different topics, we showed that the integration of visual, audio, and textual features can improve significantly over the individual use of one modality at a time.

We believe this paper made three important contributions. First, we addressed for the first time the task of tri-modal sentiment analysis, and showed that it is a feasible task that can benefit from the joint exploitation of different modalities. Second, we identified five multimodal features helpful to differentiate negative, neutral and positive sentiments: polarized words, smile, gaze, pauses, and voice pitch. Finally, we introduced a new dataset consisting of real online data, which we hope it will be useful for future research in this area.

In future work, we plan to run experiments on a significantly larger scale, so that we can also address the problem of sentiment analysis at utterance level (rather than video level, as done in this paper). We would also like to explore other domains, such as movie reviews or even political debates, which may pose additional difficulties in terms of recognizing the visual and audio features.

Acknowledgments

The authors are grateful to the three annotators who helped with the sentiment annotations. This material is based in part upon work supported by the National Science Foundation IIS award #0917170 and by the U.S. Army Research, Development, and Engineering Command. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the National Science Foundation or of the Government, and no official endorsement should be inferred.

8. REFERENCES

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

[2] J. Arguello and C. Rose. Topic segmentation of dialogue. In *HLT-NAACL Workshop on Analyzing Conversations in Text and Speech*, 2009.

[3] K. Balog, G. Mishne, and M. de Rijke. Why are they excited? identifying and explaining spikes in blog mood levels. In *Proceedings of the 11th Meeting of the European Chapter of the As sociation for Computational Linguistics (EACL-2006)*, 2006.

[4] T. Banziger and K. R. Scherer. *Introducing the geneva multimodal emotion portrayal (gemep) corpus*. Oxford University Press, 2010.

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

[6] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359, December 2008.

[7] C. Busso and S. Narayanan. Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2331–2347, November 2007.

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

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

[10] A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *LREC*, Genova, IT, 2006.

[11] G.McKeown, M. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *ICME*, pages 1079–1084, 2010.

[12] N. Godbole, M. Srinivasaiah, and S. Sekine. Large-scale sentiment analysis for news and blogs. In *International Conference on Weblogs and Social Media*, Denver, CO, 2007.

[13] J. A. Harrigan, T. E. Oxman, and R. Rosenthal. Rapport expressed through nonverbal behavior. *Journal of Nonverbal Behavior*, 9(2), 1985.

[14] M. Hu and B. Liu. Mining and summarizing customer reviews. In *SIGKDD*, Seattle, Washington, 2004.

[15] N. Kaji and M. Kitsuregawa. Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Prague, Czech Republic, 2007.

[16] R. Mihalcea, C. Banea, and J. Wiebe. Learning multilingual subjective language via cross-lingual projections. In *ACL*, Prague, Czech Republic, 2007.

[17] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *CVPR*, June 2007.

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

[19] T. Plotz and G. A. Fink. Markov models for offline handwriting recognition: a survey. *International Journal on Document Analysis and Recognition*, 12(4), 2009.

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

[21] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 1989.

[22] F. E. M. W. B. Schuller. Openear introducing the munich open-source emotion and affect recognition toolkit. In *ACII*, 2009.

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

[24] P. Stone. *General Inquirer: Computer Approach to Content Analysis*. MIT Press, 1968.

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

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

[27] J. Wiebe and R. Mihalcea. Word sense and subjectivity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006.

[28] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *CICLing-2005*, 2005.

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

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

[31] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *EMNLP*, pages 129–136, Sapporo, Japan, 2003.

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