# In the Mood for Vlog: Multimodal Inference in Conversational Social Video

Dairazalia Sanchez-Cortes, Idiap Research Institute
Shiro Kumano, NTT Comunication Science Laboratories
Kazuhiro Otsuka, NTT Comunication Science Laboratories
Daniel Gatica-Perez, Idiap Research Institute and Ecole Polytechnique Fédérale de Lausanne (EPFL)

The prevalent "share what's on your mind" paradigm of social media can be examined from the perspective of mood: short-term affective states revealed by the shared data. This view takes on new relevance given the emergence of conversational social video as a popular genre among viewers looking for entertainment and among video contributors as a channel for debate, expertise sharing, and artistic expression. From the perspective of human behavior understanding, in conversational social video both verbal and nonverbal information is conveyed by speakers and decoded by viewers. We present a systematic study of classification and ranking of mood impressions in social video, using vlogs from YouTube. Our approach considers eleven natural mood categories labeled through crowdsourcing by external observers on a diverse set of conversational vlogs. We extract a comprehensive number of nonverbal and verbal behavioral cues from the audio and video channels to characterize the mood of vloggers. Then we implement and validate vlog classification and vlog ranking tasks using supervised learning methods. Following a reliability and correlation analysis of the mood impression data, our study demonstrates that, while the problem is challenging, several mood categories can be inferred with promising performance. Furthermore, multimodal features perform consistently better than single channel features. Finally, we show that addressing mood as a ranking problem is a promising practical direction for several of the mood categories studied.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms: Human Factors

Additional Key Words and Phrases: Social video, mood, video blogs, vlogs, nonverbal behavior, verbal content

**ACM Reference Format:**
Dairazalia Sanchez-Cortes, Shiro Kumano, Kazuhiro Otsuka and Daniel Gatica-Perez, 2014. In the Mood for Vlog: Multimodal Inference in Conversational Social Video *ACM Trans. Interact. Intell. Syst.* 9, 4, Article 00 (March 2014), 25 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

## 1. INTRODUCTION

People use social media to share memories, ideas, opinions, experiences, and states of mind. In particular, social video (found on sites like YouTube, Vimeo, and Dailymotion) serves multiple purposes, including entertainment, debating, teaching and learning, and artistic expression. Social video is a major entertainment source among young audiences, and "YouTube reaches more U.S. adults aged 18-34 than any cable network" [YouTube 2014b]. The same can be said about video marketing. Furthermore,

Author's addresses: D. Sanchez-Cortes and D. Gatica-Perez, Idiap Research Institute, Centre du Parc, Rue Marconi 19, CP 592. CH-1920 Martigny, Switzerland; S. Kumano and K. Otsuka, NTT Comunication Science Laboratories, 3-1, Morinosato Wakamiya, Atsugi-shi, 243-0198, Japan.

DOI:http://dx.doi.org/10.1145/0000000.0000000

traditional media like Hollywood have been looking at how to tap into the potential of the social video medium and its huge audiences [Gillete 2014]. Conversational social video in the forms of video blogs (vlogs), video reviews, or video testimonials is a popular genre where people simultaneously share what they look like, what they think, and how they feel, in a format that is both natural and increasingly ubiquitous thanks to mobile devices.

Video blogging is a popular form of entertainment, albeit not one that older generations might be familiar with. Very popular vloggers receive millions of views, have thousands of subscribers, have achieved YouTube partner status, and get paid. According to YouTube, there are “more than a million creators from over 30 countries earning money from their YouTube videos”, and thousands of channels make six figures a year [0] The naturality and proliferation of conversational social video enable the study of human mood in this medium. Mood is defined as “a temporary state of mind or feeling” [Dictionaries 2014]. Automatic systems to analyze mood in social video could be used to search for mood trends as currently done using text from blogs or tweets [Feldman 2013; Golder and Macy 2011; Mislove et al. 2010] . They could also be used in new applications that allow for self- and community-based support, or to foster artistic expression through mood-based discovery of channels and users.

In social media, the recognition of mood from text blogs or tweets has received significant attention [Feldman 2013]. Many works have analyzed moods associated to daily life, political opinions, and population habits [De Choudhury et al. 2012; Golder and Macy 2011; Keshtkar and Inkpen 2009; Leshed and Kaye 2006; Mishne 2005; Mishne and de Rijke 2006; Mislove et al. 2010; Pang and Lee 2008], suggesting that written forms are reliable means to transmit mood. In everyday face-to-face interaction, however, we express our mood integrating speech, facial expressions, and gestures [Ekman and Friesen 2003]. The verbal and nonverbal channels inherent to co-located communication are also transmitted and perceived through remote video.

A substantial amount of research has also examined single audio or visual sources to automatically infer mood or other related variables using posed and naturalistic data [Littlewort et al. 2011; Valstar et al. 2011]. Another thread has studied the recognition of mood from multimodal cues in both scripted and realistic situations [Sebe et al. 2006; Wollmer et al. 2013]. However, the study of conversational social video from multimodal cues has not been addressed in depth, with a few exceptions [Biel and Gatica-Perez 2011, 2013; Morency et al. 2011; Wollmer et al. 2013].

The task of inferring the mood of conversational social video users can be framed in a number of ways. First, it is relevant to classify the mood of a user according to a number of intensity levels. Binary classification is a common task found in the literature related to social inference. Second, it is also useful to rank individuals according to their mood, e.g. for search purposes. As stated in [Freund et al. 2003]: “ranking models are better to fit learning problems in which scales have arbitrary values (rather than real world measures)”. For instance, a person could be labeled as looking angrier than others because the average population does not appear to be so. For certain problems, a ranking methodology could be appropriate, especially when the labels are susceptible to biased scores, as is the case of external observer annotations of mood [Mairesse et al. 2007].

In this paper, we present a systematic study on automated inference of mood in conversational social video. We study a broad set of 11 mood categories (happiness, excitement, relax, sadness, boredom, disappointment, surprise, nervousness, stress, anger, and overall mood on a diverse set of YouTube vloggers for which a rich set of nonverbal and verbal cues has been extracted. We study the vlog mood inference problem from the perspectives of classification and ranking. A preliminary version of our work was published in [Sanchez-Cortes et al. 2013]. Our contributions are as follows:

(1) We present a dataset of 264 YouTube conversational vlogs (3 minutes in average per video), which allows the study of mood categories beyond simple positive/negative polarity. The dataset, annotated via crowdsourcing, contains a variety of social video sub-genres that to our knowledge have not been collectively studied in previous work. The list includes personal experiences, entertainment, advice, reviews, and community management. The dataset reflects both the richness of the conversational social video medium and the relevance of its analysis, and highlights the key role that personal experiences play as part of the social video production and consumption cycle.
(2) We conduct reliability and correlation analyses for the crowdsourced mood annotations, finding acceptable reliability for several of the mood categories, positive correlations between some of the categories, as well as negative associations among other moods. This confirms that previous findings about mood labeling also hold for the social video setting we study here.
(3) We use state-of-the-art methods to automatically extract nonverbal features (including speaking activity and prosody from audio, and visual activity and facial expressions from video) as well as linguistic categories that have been validated in psychometric terms.
(4) We study the effect of single and combined modalities (verbal and nonverbal) on all the mood categories using supervised learning. The study shows that several categories can be discriminated in a binary classification setting, with promising results for Overall mood and Excited (69% and 68%), both statistically better than a majority class baseline. Our work shows that although multimodal features perform better than single channel features, not always all the available channels are needed to discriminate mood levels. In addition, for several mood categories, the verbal content augments the nonverbal information in the binary classification tasks.
(5) In addition to classification, we address the mood inference as a supervised ranking approach, obtaining promising results for vlog retrieval according to mood. The ranking approach is particularly interesting for mood-based search or discovery applications.

The paper is organized as follows. We discuss related work in Section 2. Our approach is summarized in Section 3. In Section 4 we describe our data. We present in Section 5 the reliability and correlation analysis of the data. Section 6 describes the nonverbal and verbal cues and the machine learning framework used in the study. We present and discuss the classification results in Section 7, ranking results in Section 8, and contrast the obtained results in both tasks in Section 9. We describe the future applicability of our findings in Section 10. We provide concluding remarks in Section 11.

## 2. RELATED WORK

Our work is related both to previous work that has examined the recognition of mood from text blogs and other social media text sources, and to work who has addressed the recognition of affective states from audio and video in face to face interactions. Each of these topics is reviewed here.

**Mood inference from text.** Studies in psychology have revealed strong connections between the words we use in written and spoken forms, and personal traits and emotional states [Mairesse et al. 2007; Pennebaker and King 1999]. It is thus not surprising that text analysis techniques have been applied in text blogs, product reviews, and social media, in the context of sentiment analysis [Feldman 2013].

Some of the first mood classification approaches focused on written blogs extracted from the LiveJournal dataset [Mishne 2005] (815k blogs, 200 words per blog on average). In this dataset, the mood labels were provided by the bloggers themselves (from a list of available moods along with an option to add new labels) when submitting the blog entries. The approaches to classify mood varied from n-grams and word statistics (including term frequency/inverse document frequency, verbs, and adjectives), to word orientation and bag-of-words [Keshtkar and Inkpen 2009; Leshed and Kaye 2006; Mishne 2005; Nguyen et al. 2010; Strapparava and Mihalcea 2007, 2008]. A number of techniques and performance measures are summarized in Table I. While it is difficult to compare the studies as mood categorization systems and classification tasks are not the same, classification accuracies have been reported to be between 24.7% and 77.6%

In the last years, a body of research has aimed at inferring mood using short text content from social media including tweets, comments, tips, etc. This task is challenging due to the brevity of the text, abbreviations, etc. As examples using tweets, the work in [Mislove et al. 2010] presented visualizations of mood fluctuations over time and space in the USA. The work presented in [Golder and Macy 2011] analyzed daily and seasonal fluctuations of mood worldwide using longitudinal data. Another work examined the potential of crowdsourcing to label mood in tweets based on the circumplex model [De Choudhury et al. 2012]. While our work also uses crowdsourced mood labels, in contrast to all the above literature, we integrate the video and audio modalities to text, and so bring in the possibility of complementing sentiment analysis techniques.

**Mood inference from audio and video.** Studies in psychology have demonstrated the relationship between affective states, including mood and emotions, and expressive human behavior. A significant body of work has also studied automated mood inference from audio and video but without specifically addressing social video. Regarding audio analysis the work in [Lee and Narayanan 2005] used acoustic features to distinguish between negative and non-negative emotions using call center data. Other affective states related to emotion have been studied in the speech community for years, with initiatives (e.g. [Schuller et al. 2011]) to compare methods appearing recently. To our knowledge, none of them have used social conversational video as we do in this paper.

Regarding visual processing, facial expressions reveal internal states [Ekman and Friesen 2003], and numerous efforts have been made to develop video-based automatic recognition systems of facial expressions [Valstar et al. 2011]. As a result, advanced facial expression analyzers are now publicly/commercially available, e.g. [Littlewort et al. 2011] and [OMRON 2007]. The analysis of spontaneous facial expressions in the wild is nowadays a key topic in affective computing. The target affective states include prototypical emotions [Valstar et al. 2011], emotional dimensions such as valence and arousal [McKeown et al. 2010], empathy [Kumano et al. 2012], pain [Littlewort et al. 2007; Lucey et al. 2012], and depression [Girard et al. 2013]. Some works have focused on the observers' impressions about the target person [Kumano et al. 2012; McKeown et al. 2010], like the present study. One recent study classified viewers' preferences for video advertisements from their smiles produced during video watching [McDuff et al. 2013]. A fundamental difference between that work and ours is that, instead of analyzing the passive behavior of observers (i.e. media consumers), we are interested in recognizing the mood of active speakers in social video (i.e., media producers).

The combination of audio and video cues to recognize affective states has also been studied in the past. A well known study in a laboratory setting reported classification of 11 emotional states using prosodic features and motion facial units from subjects displaying requested emotions [Sebe et al. 2006]. Moreover, an emotion challenge [Schuller et al. 2012] was introduced to tackle the inference of four affect dimensions (arousal, expectation, power and valence). For this challenge, a database with 24 videos is available (15 min per video). The videos contain interactions between lab participants and humans playing the role of an agent. The best scores in terms of correlation coefficient ranged between 0.174 and 0.456 against continuously annotated ground truth.

To our knowledge, the closest works to ours are [Morency et al. 2011; Wollmer et al. 2013]. In [Morency et al. 2011], 47 videos from YouTube where people reviewed products were studied. Each video was normalized to 30-second duration, and the extracted features included gaze, smile, word polarity, pause, and pitch. The videos were manually labeled as negative, neutral or positive. While single modalities showed low performance in terms of F-measure, additional experiments using multimodal features showed an increase of performance up to 55.3%.

The work in [Wollmer et al. 2013] presented binary classification of sentiment polarity using 370 videos from movie reviews extracted from YouTube and ExpoTV. The labeling was performed by two annotators for the YouTube videos and a single annotator for the ExpoTV videos. The paper presented a comparison of multimodal cues including audio, visual (facial expression, gaze, and smile) and linguistic features (from manual transcriptions and ASR), reporting up to 73% weighted F1-measure. A key difference between [Morency et al. 2011; Wollmer et al. 2013] and our work is that we are interested in studying social video with a wider diversity of topics and not only movie or product reviews. Moreover, we study and report performance on 10 mood categories plus overall mood (i.e., overall judgment of positive/negative mood), while [Morency et al. 2011; Wollmer et al. 2013] only focus in the latter category.

In [Sanchez-Cortes et al. 2013], we presented a preliminary version of this work. In this paper, we extend our previous study in three ways. First, we present an in-depth analysis of our vlog dataset from the perspective of topics and correlation analysis of mood annotations. Second, we study mood inference from the perspective of ranking in addition to classification. Finally, we study the correlation between classification and ranking methods.

Table I. Related Work on Mood. Modalities: T=Text, A=Audio, V=Visual, L=Verbal. Studied Tasks: B=Binary, M=Multiclass, C=Continous (Regression u other). Performance measures: Acc=Accuracy, F1=F1-Score, AUC=Area Under the Curve, CC=Correlation Coefficient.

| Reference | Data Modality | Data Source | Mood Categories | Task | Performance Measure | Reported Performance |
|---|---|---|---|---|---|---|
| Leshed, 2006 | T | Blogs LiveJournal | 50 top moods | B | Acc | 74% |
| Keshtkar, 2009 | T | Blogs<br>LiveJournal | (1) 132 moods and (2) 15 moods<br>e.g happy, sad, angry | M | (1) Acc<br>(2) Acc | 24.73%<br>63.5% |
| Nguyen,2010 | T | Blogs WSM09 (1)<br>IR05 (2) | Happy, sad, angry | B | F1 (1)<br>F1 (2) | 0.697-0.774<br>0.709-0.788 |
| Mishne,2006 | T | Blogs<br>LiveJournal | 132 top moods<br>Happy, sad, angry, etc | C | CC | 0.95 (Happy)<br>0.79 (Angry) |
| Strapparava,2007 | T | News headlines | 6 moods | C | Acc | 93.6% (Angry) |
| Strapparava,2008 | T | News headlines | 6 moods | C | F1 | 0.30 (Sad) |
| Mislove,2010 | T | Tweets | Happy | n.a. | n.a. | n.a. |
| Golder,2011 | T | Tweets | Affect (-,+) | n.a. | n.a. | n.a. |
| Chouhdury,2012 | T | Tweets | 200 moods | n.a. | n.a. | n.a. |
| Nicolaou,2011 | A | Lab sessions | Valence (-),Arousal (+) | C | RMSE | 0.25(-), 0.26(+) |
| MinLee,2005 | A,L | Call center | Positive, negative | B | Error | 14.1(M),13.8(F) |
| Mckeown,2010 | V | Lab sessions | Valence (-),Arousal (+) | n.a. | n.a. | n.a. |
| Valstar,2011 | V | Emotion portrayals | 12 Action Units (AU), 5 emotions (E) | M | F1, Acc | 0.45 (AU), 0.56 (E) |
| Mcduff,2013 | V | Response to ads | Liking (L), desire to watch again (D) | B | AUC | 0.8 (L),0.78 (D) |
| Sebe,2006 | A,V | Scripted Videos | 11 affect categories | B | Acc | 90% |
| Schuller,2012 | A,V | Lab sessions | Arousal, expectation, power, valence | C | CC | 0.174-0.456 |
| Morency,2011 | A,V,L | YouTube review Prod. | Positive,negative and neutral | M | Acc | 55.3% |
| Wollmer,2013 | A,V,L | Movie review | Positive and negative | B | F1 (weighted) | 73% |
| Sanchez-Cortes,2013 | A,V,L | YouTube vlogs | 11 moods | B | AUC | 0.74 (Excited) |

## 3. OVERVIEW OF OUR APPROACH

Figure 1 presents our approach. We use 264 vlogs from YouTube with mood annotations obtained via crowdsourcing, where each vlog is annotated for 11 natural mood categories. The vlogs contain a person that discusses personal experiences, expresses opinions, and interacts with their audiences. We performed an analysis on the data to verify the quality of our ground-truth labels, reviewed the diversity of our data in terms of topic and performed correlation analysis on the annotations.

From each vlog, we systematically extract a number of nonverbal and verbal cues that allow multimodal analysis. With the multimodal features we then use a classifier. We propose single features and fusion of features to investigate the discriminative value of each channel. We use feature concatenation for fusion. For each mood category, we define a binary classification task to discriminate vlogs as being above or below the median of the population. Moreover, we apply ranking methods to provide a list in which the top positions are the most representative vlogs for a given mood. Finally, we analyze the outputs of the proposed methods and analyze their correlation performance.

The nonverbal features include **audio cues**, i.e., acoustic features including pitch, energy, speaking rate, formants and bandwidths computed from the audio channel; **visual features** that capture looking activity, pose cues and visual activity, and facial expression cues; and verbal cues from which we computed word categories using Linguistic Inquiry Word Count (LIWC) from manual transcriptions. We describe the feature extraction process in Sections 6.1 and 6.2.

Regarding classification, we have 11 mood categories as stated earlier. We divided the samples per mood using the median value from the mood labels, and applied 10-fold cross-validation, where train and test sets are disjoint. The features are normalized using z-normalization and passed to the binary classifier (e.g., Happy and Non-Happy), in this case Random Forest (RF). We first study features from single modality cues, and then we perform feature fusion.

Regarding ranking, we trained a learner per mood. For every pair of vector features, we use the mood ground-truth order rank to generate a learning vector (as long as one of the instances is ranked higher than another). We also applied 10-fold cross-validation, and we use the same normalized features used by the classification approach.

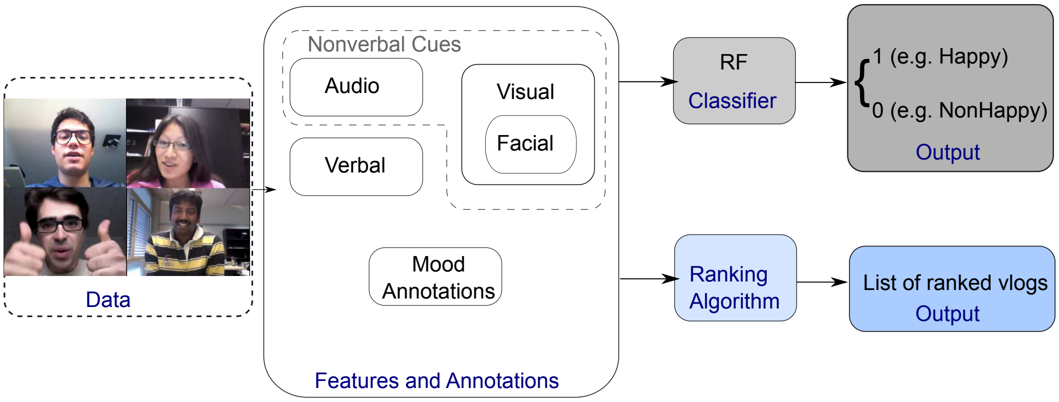


Fig. 1. Overview of our approach.

## 4. DATA

We used the dataset of YouTube conversational social video by Biel and Gatica-Perez [Biel and Gatica-Perez 2013]. This data includes 264 vlogs, each one featuring one single vlogger talking in English. The collection had no restriction in terms of the topics addressed by the vloggers or the recording setting, so the dataset is quite diverse with respect to the content and the audio and visual quality of the videos. The typical vlog is recorded indoors with a commercial webcam, lasts about three minutes, and features the head and shoulders of the vlogger. Figure 2 shows an example of the vlog corpus including transcription and some of the automatically extracted features.

The dataset also includes annotations of mood and demographic impressions that were collected from people watching vlogs in Mechanical Turk [Biel and Gatica-Perez 2012]. The reason to use non-experts in the annotations is supported by the findings reported in [Ekman and Friesen 2003] and [Snow et al. 2008], which affirm that untrained observers can accurately judge spontaneous and natural emotions. Moreover, one of the advantages of labeling mood via crowdsourcing is that the annotators watch the video in ecologically valid conditions, i.e., watching them directly on YouTube. Concerning demographics, approximately 70% of the vloggers were labeled as below 24 years old, and around 80% of the population was reported as Caucasian. With respect to gender, it is mostly balanced: 53% females and 47% males. Clearly our sample is not a fair sample of the world population, but reflects the statistics of the YouTube, English-speaking video blogger community.

For each vlog, five Mechanical Turk workers annotated the ten different moods, as well as overall mood (overall judgment of positive/negative mood) using a 7-point likert scale. The list of moods came from the Livejournal text blogging platform, and from here a subset of mood adjectives was selected, considered as possibly manifesting in vlogs. The list covers ten different affective states of diverse arousal and valence, and one item contains the overall mood valence. From the 11 mood categories presented in this paper, six are the same as reported in [Sebe et al. 2006]. The choice of five workers for the annotation task is supported by the findings of Snow et al. [Snow et al. 2008] who empirically found that "for an affect recognition task we find that we require an average of 4 non-expert labels per item in order to emulate expert-level quality". Note however that the task in [Snow et al. 2008] and ours is not the same, since we use different data sources, i.e. vlogs rather than news text headlines.

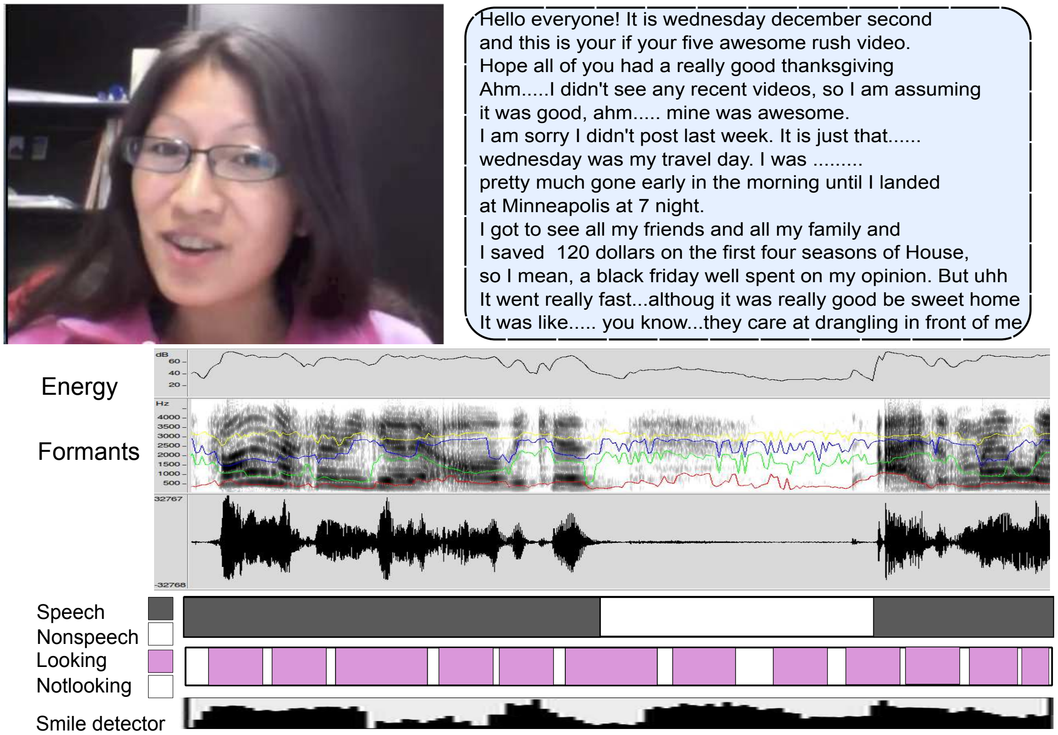
We complemented this dataset with the manual transcriptions of vlogs, which was performed by professionals. The transcriptions have in average 625 words per vlog. For comparison, the average number of words using blogs in related works include [Leshed and Kaye 2006] with 168 words, [Keshtkar and Inkpen 2009] with 200 words, and [Mishne and de Rijke 2006] with 140 words per blog.

## 5. DATA ANALYSIS

### 5.1. Reliability analysis

As measure of reliability for the annotations, we use the Intraclass Correlation Coefficient measure $ICC(1,k)$, which is a standard measure used in psychology. ICC is a measure of similarity that assesses consistency of quantitative measurements made by different observers [Koch 1982]. The ICC is the proportion of the total variance within our data that is explained by the variance between annotators. ICC(1,k) means that each vlog is assessed by a different set of randomly selected annotators, and the reliability is calculated by taking an average of the k annotators' measurements ($k = 5$).

$$ICC(1,k) = \frac{BMS - WMS}{BMS} \tag{1}$$



Hello everyone! It is wednesday december second
and this is your if your five awesome rush video.
Hope all of you had a really good thanksgiving
Ahm.....I didn't see any recent videos, so I am assuming
it was good, ahm..... mine was awesome.
I am sorry I didn't post last week. It is just that......
wednesday was my travel day. I was .........
pretty much gone early in the morning until I landed
at Minneapolis at 7 night.
I got to see all my friends and all my family and
I saved 120 dollars on the first four seasons of House,
so I mean, a black friday well spent on my opinion. But uhh
It went really fast...althoug it was really good be sweet home
It was like..... you know...they care at drangling in front of me

Fig. 2. Example of the vlog corpus, including transcription and Extracted features.

where BMS=between annotations mean square, WMS=within annotations mean square. ICC varies between 0 and 1. When ICC approaches 1, this indicates very high agreement between annotators. The judgments are averaged across annotators (used as ground-truth in our paper), and are reliable with the following intra-class correlations: Overall mood (.75), Happy (.76), Excited (.74), Angry (.67), Disappointed (.61), Sad (.58), Relaxed (.54), Bored (.52), Stressed (.50), Surprised (.48), Nervous (.25). These ICC values show that high arousal moods such as Excited, Happy, or Angry are easier to judge by annotators, a result that might be explained by these moods manifesting themselves more explicitly in the behavior of vloggers.

### 5.2. Vlog Categories

To assess the diversity of topics in the YouTube dataset, we performed a manual annotation of video categories that describe the video content, with one annotator (the first author of this paper). The list of the categories was formed considering the standard 19 YouTube channel categories [YouTube 2014a], and adding categories that are not included in the list but that are relevant to the vlog context. We chose six YouTube categories which include: (1) Entertainment (including the categories Comedy, Film and Entertainment, Animation, Music, and Sports), (2) News and Politics, (3) Non-profits and Activism, (4) HowTo and Style (including the categories How to and Do it yourself, and Beauty and Fashion), (5) SciTech and Education (which includes the categories Technology, and Science and Education), and (6) Cooking and Health. In addition, we defined additional categories relevant to conversational vlogs: (1) Personal Experience (that includes events of daily life), (2) Advice (giving advice on a informal topic), (3) Channel Managing (i.e., promoting a YouTube channel or other social media like Twitter, Facebook, or Picasa, and replying to questions and comments), (4) Product Review (where products are movies, restaurants, museums, etc.), and (5) Religion and Ideology. The annotation procedure was as follows: the annotator first watched the vlog and then chose one or two labels that best described the content, choosing freely among all categories.

Table II. Distribution of Categories in the YouTube dataset. The first 6 categories correspond to the predefined YouTube categories. The 5 last categories were defined through manual annotation.

| **Category** | **Percentage** |
|---|---:|
| Entertainment | 7.5% |
| News and politics | 3.7% |
| Non-profits and Activism | 3.7% |
| HowTo and Style | 3.2% |
| ScienceTechnology and Education | 2.0% |
| Cooking and health | 1.4% |
| Personal experience | 54.9% |
| Advice | 8.3% |
| Channel Managing | 7.5% |
| Product review | 6.0% |
| Religion and ideology | 1.7% |

Table II presents the distribution of the manually annotated categories. The top YouTube category (Entertainment) represents 7.5% of the labels in the data, followed by the categories News and Politics, and Non-profits and Activism with 3.7%. On the other hand, the Personal Experience label represents 54.9% of the labels in the data, followed by Advice with 8.3%, and Channel Managing with 7.5%. Regarding the number of labels per vlog, 68.2% of the vlogs were annotated with a single category, and 31.8% were given 2 category labels. It is worth to mention that Personal experience was the most common category when 2 labels were needed. The large amount of vlogs in the Personal Experience category also suggest that moods in the dataset are naturalistic.

### 5.3. Correlation analysis

We performed a Pearson correlation analysis to understand which mood impressions could appear together. Pearson correlation coefficient is a measure of the strength of the linear relationship between two variables, defined as

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, \tag{2}$$

where $cov$ is the covariance, $\sigma$ is the standard deviation, and $X$, $Y$ refer to each of the 10 moods. The correlation is computed using the averaged values for each of the 10 moods. In Table III, we present correlation values that are statistically significant with $p < 0.005$.

As we can observe, there are strong correlations among some moods. We discuss them in descending order according to their ICC reliability:

— Happy is strongly and positively correlated with Excited (0.82), and although weaker, it has a positive correlation with Relaxed. As expected, Happy has negative correlations with Disappointed (-0.72), Sad (-0.69), Stressed (-0.64) and Angry (-0.60). Also, it has negative significant correlations with Bored and Nervous. No significant correlation was found with Surprised.

Table III. Pearson correlation among mood (N=264, $p < 0.005$). Categories ordered according to their ICC reliability value (see Section 4).

| Pearson-Corr | Excited | Angry | Disappointed | Sad | Relaxed | Bored | Stressed | Surprised | Nervous |
|---|---|---|---|---|---|---|---|---|---|
| Happy | 0.82 | -0.60 | -0.72 | -0.69 | 0.29 | -0.54 | -0.64 | - | -0.40 |
| Excited | | -0.28 | -0.57 | -0.68 | - | -0.63 | -0.56 | 0.35 | -0.36 |
| Angry | | | 0.68 | 0.44 | -0.45 | 0.28 | 0.53 | 0.24 | 0.32 |
| Disappointed | | | | 0.79 | -0.31 | 0.48 | 0.75 | 0.19 | 0.47 |
| Sad | | | | | -0.18 | 0.61 | 0.78 | - | 0.51 |
| Relaxed | | | | | | - | -0.34 | -0.24 | -0.29 |
| Bored | | | | | | | 0.48 | -0.22 | 0.38 |
| Stressed | | | | | | | | - | 0.71 |
| Surprised | | | | | | | | | - |
| Nervous | | | | | | | | | |

- Excited shows moderately positive correlation with Surprised (0.35). Correlations are negative for the other moods (from -0.68 to -0.28).
- Angry has positive strong correlation with Disappointed (0.68), followed by Stressed, Sad, Nervous, Bored and Surprised. Angry has also a significant negative correlation with Relaxed (-0.45).
- For Disappointed, we can observe strong positive correlations with Sad and Stressed (0.79 and 0.75 respectively), followed by Bored and Nervous. There is also a negative significant correlation with Relaxed (-0.31).
- For Sad, we can observe a strong positive correlation with Stressed (0.78), followed by Bored (0.61) and Nervous (0.51), and negative weak correlation with Relaxed (-0.18).
- For Relaxed, we can observe weak to moderate negative correlations with Stressed, Nervous and Surprised (-0.34, -0.29 and -0.24 respectively).
- For Bored there is a positive significant correlation with Stressed (0.48), followed by Nervous (0.38). Moreover, there is a weak negative correlation with Surprised (-0.22).
- Finally, there is a strong correlation between Stressed and Nervous (0.71).

Overall, the correlation matrix shows connections that were expected for several moods, some of which have been reported in previous literature.

## 6. AUTOMATIC MOOD INFERENCE

We integrated several audio processing, computer vision and text analysis technologies to characterize vloggers' nonverbal and verbal behavior. We first describe the methods used to compute nonverbal cues from audio and video, then explain the analysis technique used to characterize verbal content. Finally, we give details about the classification and ranking supervised methods.

### 6.1. Nonverbal Cues

We investigate three nonverbal behavioral sources that have been documented in nonverbal communication research as conveying emotional information [Knapp and Hall 2008]: vocal cues, visual activity, and facial expressions.

*6.1.1. Audio nonverbal cues.* Voice is a primary channel for expressing emotion [Knapp and Hall 2008]. Research has shown that emotion perception depends on changes in pitch, volume, and speaking rate [Scherer 2003], and has repeatedly showed that automatically extracted prosodic cues are useful to capture personal and emotional information [Lee and Narayanan 2005; Sebe et al. 2006].

We extracted **prosodic cues** that estimate the pitch, energy, and speaking rate of vloggers. First, we processed the audio channel of vlogs using PRAAT [Boersma 2002] to generate frame-by-frame estimates of these and other related signals (e.g. the second and third formants and their bandwidth). Second, we aggregated features across the whole video duration by computing the mean, median, mean-scaled standard deviation, maximum, minimum, and entropy. In total, we computed 98 prosodic cues.

*6.1.2. Visual activity nonverbal cues.* Gesture, gaze, posture, and movement can reveal cognitive and affective states. The extraction of these nonverbal cues in social video is challenging due to the variety of content available, but has nevertheless been addressed to build computational models of vlogger personality [Biel and Gatica-Perez 2013].

We extracted three types of visual nonverbal cues. First, we extracted **looking activity cues** (cues related to gaze) obtained from looking-non-looking segmentations including the time looking at the camera, the average duration of looking segments, and the number of looking turns. These segmentations were produced following a method based on a frontal face detector [Biel and Gatica-Perez 2011]. Second, we used the position and size of detected faces to compute **pose cues** such as the proximity to the camera and the horizontal and vertical framing of the vlogger (i.e., the position of the vlogger with respect to the center of the frame). Finally, we characterized the **visual activity** of vloggers through the computation of weighted motion energy images (wMEI). wMEIs are gray scale images that measure the accumulated motion through the whole video (one single image is generated per video, where brighter pixels correspond to regions with higher motion). For the frontal face detection, we used implemented Haar-like features on OpenCV in order to scan faces as small as 20x20 pixels [Bradski and Kaehler 2008]. From the visual nonverbal cues, we computed several features such as the entropy, mean, median, and the vertical and horizontal center of mass.

In addition to the visual activity features, we also extracted a few **multimodal cues** generated from looking/not-looking and speech/non-speech segmentations. In particular, we computed the looking-while-speaking time (L&S), the time looking-while-not-speaking (L&NS), and the multimodal ratio (L&S/L&NS), which capture joint patterns of speech and gaze. The total number visual and multimodal cues sums up to 31.

*6.1.3. Facial expressions.* Facial expressions are important cues in human perception [Knapp and Hall 2008], accounting for personality traits [Ambady and Rosenthal 1992], as well as cognitive and psychological states [Ekman and Friesen 2003]. Today, real-time facial analysis can be addressed with tools such as the Computer Expression Recognition Toolbox (CERT) [Littlewort et al. 2011]. Though these technologies were developed for videos without speech, research has also shown that automatic facial expression cues derived from CERT can be used to predict vlogger personality [Biel et al. 2012]. In our research group, we evaluated the accuracy of this module to recognize facial expressions on a 1600-vlog frame set that was annotated with respect to facial expressions using a crowdsourcing approach. The results on a task where a single dominant expression was recognized show that Joy is identified in 80% of the cases by both CERT and human annotators, Surprise in 33%, and Disgust, Anger and Sad in 22%.

We followed the approach used in [Biel et al. 2012] to aggregate the frame-by-frame outputs of CERT. CERT detects frontal faces and codes each frame with respect to 40 dimensions, including expressions of anger, disgust, fear, joy, sadness, surprise, contempt, a measure of head pose, and 30 facial action units from the Facial Action Coding System.

First, we converted frame-by-frame estimates to a binary segmentation that divides each expression signal into active/inactive regions, and then we computed features such as the duration of active time and the number of active turns. Active/inactive segmentations generate 27 facial expression cues.

### 6.2. Verbal Cues

Social psychology research has shown that the words people use reflect information about psychological constructs [Pennebaker et al. 2001]. Text can be analyzed using tools such as the Linguistic Inquiry Word Count (LIWC), which categorizes words into linguistic and paralinguistic categories that have been validated in psychometric terms. This tool has been previously applied to analyze essays and text blogs [1].

We use verbal content to infer mood through the analysis of manual transcriptions of vlogs. Each transcript was processed with LIWC to breakdown the word category usage based on relative word occurrences. The LIWC dictionary is composed of almost 4,500 words [Pennebaker et al. 2001]. Each word belongs to one or more word categories. For example, the word "agree" is part of three word categories: *affect*, *posemo*, and *assent*. So, whenever the word "agree" is found, the scores in these categories will be incremented. More details on the categories and the dictionary can be found in [LIWC 2007]. The word categories generated by LIWC are used as features, by representing each vlog by a 62-dimension vector, where each dimension corresponds to the count for each LIWC category.

We also explored the performance of unigrams. Following the methodology proposed in [Biel et al. 2013], each transcript was preprocessed by removing punctuations and discarding words with low frequency (less than ten documents). Stop words were not removed in order to have a fair comparison with the word LIWC categories. Then, the unigrams were generated followed by the computation of term frequency$-$inverse document frequency ($tf \cdot idf$). The experiments were performed considering the top 200 unigrams, the respective distribution is shown in Figure 3.

**Automatic Speech Recognition (ASR).** The performance of ASR in the mood prediction was also explored. We used an in-house ASR system that employs a lexicon of 50,000 words and a 4-gram language model [Hain et al. 2012]. The performance of the ASR system in our YouTube dataset reached 62.4% word error rate (WER) (see [Biel et al. 2013] for more details). For further analysis, each automatic transcription was also processed with LIWC.

### 6.3. Mood Classification

To classify mood in vlogs, we use Random Forest Regression as it does not tend to overfit (it uses out-of-bag samples to estimate the generalization error), it is fast to build (as it grows trees in parallel), it is robust to outliers, it can handle data from mixed types, and it performs automatic selection of features [Breiman 2001].

We train a supervised regressor per mood (k={happy, excited, ...}) using single and multimodal cues, where the input vector contains the respective set of features ($f$). In the test phase, the outputs from the learner are thresholded (using the median value) to perform two-class classification per mood.

$$Mood_k^f(vlog_i) = \begin{cases} 1 \text{ if } y(vlog_i) \geq Median_k; \\ 0 \text{ if } y(vlog_i) < Median_k. \end{cases}$$

Where $mood_k^f$ means the label assigned to $vlog_i$, tested with mood classifier $k$ (k={happy, excited, ...}) given features $f$. The output of the classifier $y(vlog_i)$ is then thresholded using the median value of the mood $k$. Later on, we estimate the significance of the accuracy (at 95% confidence level) using a two-tailed standard binomial significance test with $z = N(0, 1)$ [Lowry 1998] with respect to the baseline. The baseline per mood corresponds to majority class performance. Given that several values

[1] http://www.liwc.net

are equal to the median, the baseline is not exactly 50% (as it would be expected in a random binary task).

### 6.4. Mood Ranking

Classification tasks are hard decision methods that could be affected in terms of performance if several samples lie on the borderline class. Considering this, and the fact that mood annotations are susceptible to personal ratings and interpersonal differences, we applied ranking methods to the mood inference problem. With this task, the goal is to correctly order the vlogs according to their rank, rather than assigning them to a binary category. Ranking is a naturally useful task in search and discovery.

Ranking methods can be seen as a classification problem of order of pairs. The classification method projects the pairs and sorts them according to the projection. In other words, each pair provides the information of which instance should be ranked higher or lower with respect to the other, and the algorithm tries to minimize the number or misordered pairs. In [Mairesse et al. 2007], a ranking approach was applied to personality inference using acoustic cues, and the reported results were significantly better than the baseline. We follow this approach for mood inference as a ranking problem. For the ranking algorithm, we applied the SVM ranking model denoted as follows:

$$min_w \sum_{i,j} max(0, < w, x_i - x_j > * eval(y_i \leq y_j)) + \lambda||w||^2 \tag{3}$$

where the weight vector $w$ corresponds to the ranking function. The training consists of pairs of feature vectors $x_i$ and $x_j$, and mood scores $y_i$ and $y_j$ that tell which vector should be ranked on top, i.e., $eval(y_i \leq y_j)$ =1 if the inequality is true and 0

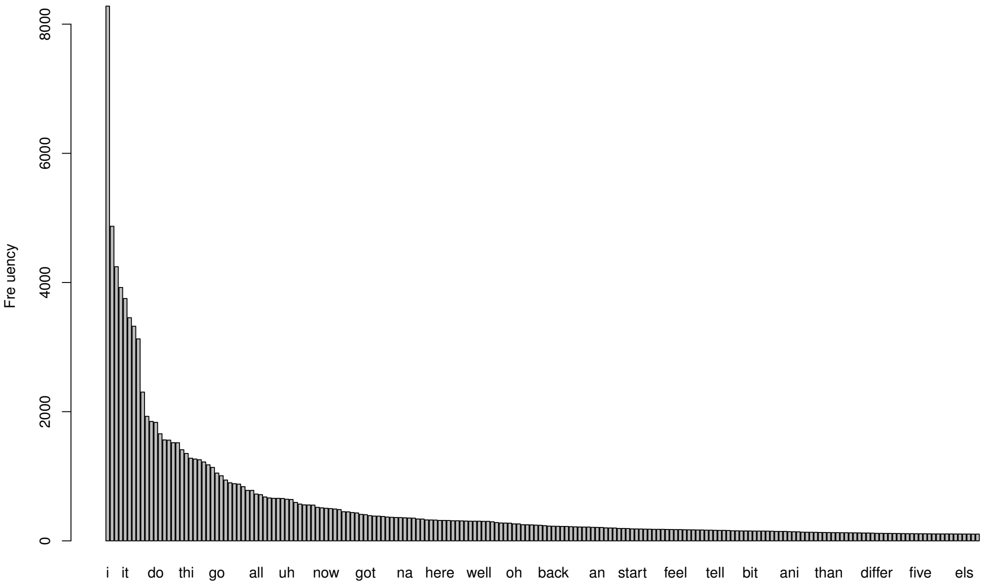


Fig. 3. Top 200 Unigrams from the vlog manual transcriptions.

otherwise. In the training phase, we performed optimization of parameters using the NRBM (Non-convex Regularized Bundle Method) method [Do and Artières 2012] on a 10-fold cross-validation approach.

For the testing phase, we estimated three well-known performance measures in information retrieval: average precision (AP), recall, and F1. For average precision, vlogs retrieved at the top of a list are more important than vlogs towards the bottom: this measure assigns more weight to the errors made at the top of a ranking. We report the average precision, recall, and F1 at top 10, averaged over the 10 folds.

The ground truth for the ranking algorithm corresponds to the sorted lists per mood, considering the values from the averaged mood annotations described in Section 4. Since we applied 10 fold cross-validation, the ground truth corresponds to the sorted vlogs list in each fold.

In addition, we estimated the Kendall rank correlation coefficient [Kendall 1975] to measure the similarity between the ground truth rankings and the rankings estimated by the algorithm. This measure takes into account ordered pairs and penalizes disordered pairs, such that a perfect ranking will provide high correlation (i.e., 1), and a negative correlation means that the ranks are reversed. The Kendall rank correlation coefficient ($\tau$) is defined as follows:

$$\tau = \frac{C - D}{\sqrt{\frac{1}{2}n(n-1) - T}\sqrt{\frac{1}{2}n(n-1) - U}}, \tag{4}$$

where $C$ is the number of concordant pairs (i.e., if both $g_i < g_j$ and $r_i < r_j$, or $g_i > g_j$ and $r_i > r_j$), $g_i$ and $g_j$ are elements of the ground truth list, and $r_i$ and $r_j$ are elements of the list corresponding to the ranking algorithm, $D$ is the number of discordant pairs, $n$ is the number of samples equal to the number of samples in each test fold, $\frac{1}{2}n(n-1)$ represents the total number of ordered pairs, and $T$ and $U$ are the number of ties in the compared lists. A tie is defined as a pair of samples with the same rank, i.e., both samples have the same averaged score (for the ground truth, $T$) or the same estimated ranking score (from the ranking algorithm, $U$).

## 7. MOOD CLASSIFICATION RESULTS

In this section, we present the results for the classification task. The results are organized per cue modality, followed by a discussion about the best results obtained for each mood. Although the results are discussed by modality, we grouped the results per mood in Figure 4 for better intuition of which mood performs better with respect to modalities or combination of features. In the figure, the solid blue line represents the majority class baseline performance (note that this is around but not exactly 50% as discussed earlier; and the red dashed line corresponds to performance that is statistically better than the baseline at 95% confidence interval.

### 7.1. Audio Nonverbal Cues (A)

For Audio features (A) as single modality, the performance for 9 moods is not statistically better than the baseline. In Figure 4, we only observe significant performance improvement for Excited and Bored at 61.9% and 60.6% respectively.

### 7.2. Visual Activity (V) and Facial Expression (F) Cues

The visual activity channel (V) includes gaze, posture, motion and gaze and speaking patterns, described in Section 6.1. From Figure 4, for Excited we observe that these cues perform significantly better than the baseline (65.3%). This could be explained by the fact that highly excited vloggers exhibit high motion in their videos. Moreover, we can observe that Visual activity cues can infer three additional moods including Disappointed (62.6%), Sad (59.6%) and Bored (61.0%).

Facial expressions (F) as single cue can infer Happy and Excited moods statistically better than the baseline. For Happy, we obtain 62.4%; perhaps explained by the accurate detection of smiles from frontal faces in the video. For Excitement, we obtain 60.3%, possibly due to the accurate CERT detection of basic expressions of joy and smiles. We also can observe statistically significant accuracy for Overall mood and Bored, 58.4% and 61.4% respectively.

### 7.3. Verbal Cues (L)

For the verbal cues, we first performed a comparison between three methods: LIWC from manual transcriptions, LIWC from ASR, and unigrams. The best performance from ASR/LIWC is for Nervous 59.0%, followed by Bored 57.7%. For Disappointed, Overall, Happy, and Sad the performance is 56.5%, 56.5%, 54.0% and 51.9% respectively.

Moreover, the best performance for unigrams is 59.8% for Sad. The performance for Excited, Disappointed, Happy and Overall mood is 59.0%, 59.0%, 56.9% and 56.1% respectively.

As the problem of automatic speech recognition in unconstrained domains like YouTube is still an open issue [Hinton et al. 2012], and these results are not statistically better than the baseline for many cases, we decided to continue the analysis only considering verbal cues derived from the manual transcriptions. Similarly, we did not observe statistically significant improvements by using unigrams, as compared

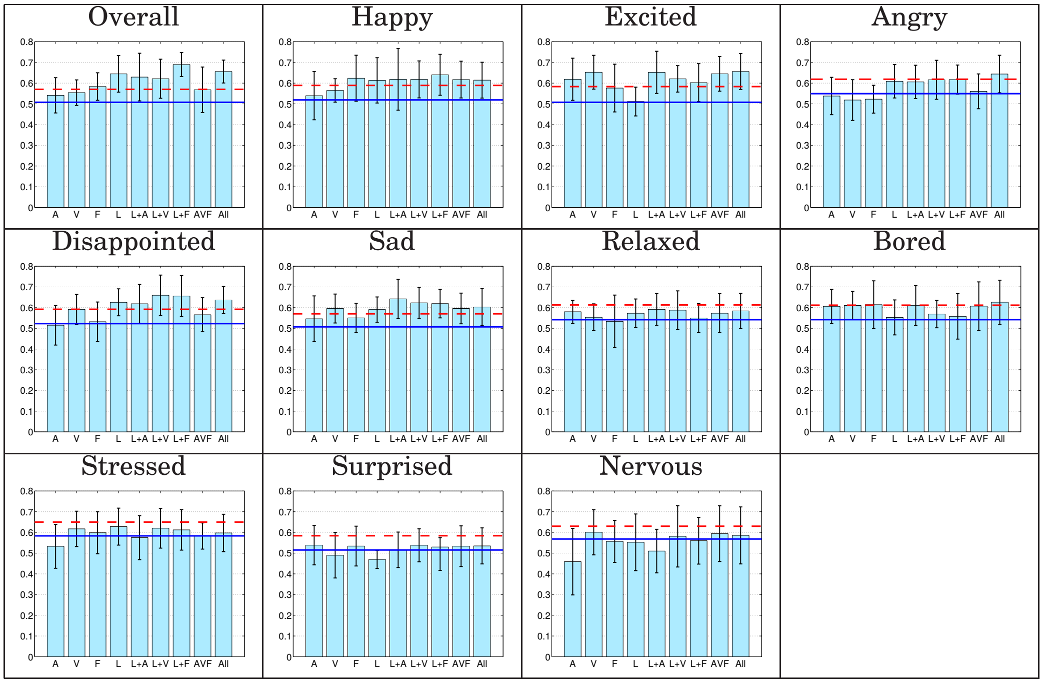


Fig. 4. Mood Classification Accuracy comparison using **RF**. Moods are ordered according to their ICC reliability value (see Section 4). A: Audio, V: Visual, F: Facial, L: Verbal, L+A: Verbal and Audio, L+V: Verbal and Visual, L+F: Verbal and Facial, AVF: Audio, Visual and Facial, All: All features. Blue solid line: Baseline method (Majority class), Red slashed line: Significantly better than baseline at 95% confidence interval.

with performance from LIWC categories from transcriptions. We thus continue the discussion of this Section focusing on the manual/LIWC method.

The LIWC word categories, show significantly better performance than the baseline for the Overall mood (64.5%). Happy (61.3%), Disappointed (59.1%) and Sad (59.1%).

We performed an analysis per mood to review the most relevant word categories per mood. For the analysis we obtain the importance of each LIWC category, using the importance component of the random forest per fold, furthermore we accumulate the 10 folds to obtain the overall importance per word category.

Table IV. Top 10 Relevant word categories for Happy, Excited, Disappointed, Sad and Nervous mood.

| Happy | | Excited | | Disappointed | | Sad | | Nervous | |
|---|---|---|---|---|---|---|---|---|---|
| Category | Relevance | Category | Relevance | Category | Relevance | Category | Relevance | Category | Relevance |
| health | 82.6 | health | 81.6 | health | 88.3 | health | 137.8 | health | 41.7 |
| swear | 81.1 | nonfl | 47.2 | negemo | 83.6 | Dic | 71.7 | Dic | 30.4 |
| posemo | 75.4 | posemo | 45.3 | swear | 77.8 | quant | 68.0 | funct | 20.8 |
| anger | 44.6 | assent | 31.1 | posemo | 60.5 | assent | 30.1 | swear | 19.3 |
| nonfl | 43.5 | funct | 28.4 | quant | 52.2 | funct | 25.2 | anx | 17.1 |
| adverb | 43.3 | tentat | 27.5 | Dic | 34.0 | humans | 24.8 | negemo | 16.7 |
| quant | 37.7 | affect | 25.8 | anger | 31.6 | bio | 22.1 | WPS | 16.5 |
| negemo | 33.0 | Dic | 24.8 | social | 26.6 | social | 20.0 | Sixltr | 14.7 |
| tentat | 24.8 | anger | 24.4 | bio | 24.3 | Sixltr | 18.6 | affect | 13.7 |
| bio | 21.7 | insight | 18.7 | adverb | 23.0 | insight | 17.7 | posemo | 11.9 |

Table IV shows the relevance of the top 10 word categories for five of the moods. As we can observe for Happy mood, the top relevant LIWC word categories are *health*, *swear* and *posemo* (positive emotion), followed by *anger*, *nonfl* (non fluencies), *adverb* (adverbs), *quant* (quantifiers) and *negemo* (negative emotion). For Excited, the top relevant categories include *health*, *nonfl*, *posemo*, *assent* (assent, e.g. agree, OK, yes), *funct* (total function words), and *tentat* (e.g. any, depend, if, some). For Disappointed, the top relevant categories are *health*, *negemo* and *swear*, followed by the categories *posemo*, *quant*, *Dic* and *anger*. For Sad mood, the top relevant categories are *health*, *quant*, *Dic* (dictionary words), followed by the categories *assent*, *funct*, *humans* (e.g. adult, baby, boy) and *bio* (biological processes). It is not surprising that Sad mood can be inferred if the verbal content reveal high percentages on word categories like *health* (e.g., alive, cancer,flu, headache, ill, life, pain, sick, overweight), *quantifiers* (e.g., a lot, anymore, less) and *humans*. For Nervous, top relevant categories include *health*, *Dic*, *funct*, *swear* and *anx*

During the manual annotation of categories described in Section 5.2, we observed the prominence of the category *health* used in three contexts. First, several vloggers apologize themselves during the first few seconds, for their vlogging absence due to sickness. Second, several vloggers provide periodic updates on their health issues like overweight, cigarette smoking, etc. Third, vloggers promote raising money to fund research for chronic diseases like cancer, Alzheimer, etc.

### 7.4. Multimodal Cues

For Overall mood, Happy and Disappointed, the best multimodal combination is with Verbal and Facial Expression Cues (**L+F**). As we can observe in Figure 4, Overall mood and Happy reach accuracies of 68.98% and 64.0% respectively. We also observe that the best multimodal combination using Audio, Visual and Facial Cues (**AVF**, i.e., only nonverbal cues), performs the best for Excited (68.3%). For Disappointed, the best multimodal combination is using Verbal and Visual Cues (**L+V**) with 65.96%. For Angry and Bored, **All** the features (Audio, Visual, Facial and Verbal Cues) are needed to reach the best performance (64.4 and 64.1% respectively).

### 7.5. Discussion of Best Results

Table V shows the summary of best accuracy achieved per mood. Moods are ordered according to their ICC reliability. Moods whose ICC $> 0.5$ are above the line and for ICC $\leq 0.5$ are under the line. Note that we only include results for which the performance is statistically better than the baseline. This shows that four moods could not be classified better than majority class (Relaxed, Stressed, Surprised and Nervous, see empty entries in Table V). The overall mood task resulted in the highest performance (69% accuracy). Two observations are the following. First, for all moods it was a combination of features (although not necessarily the same ones) what produced the best performance. Second, we do not observe any clear pattern between performance and reliability for moods with ICC $> 0.5$. This means that the reliable moods tend to produce similar performance than less reliable ones (which correspond to noisier tasks). That said, the results for the least reliable moods (Stressed, Surprised, Nervous, ICC $\leq 0.5$) are not statistically significant.

Table V. Best classification results per mood. Moods are ordered according to their ICC reliability. All non-empty entries are Statistically better than baseline at 95% confidence interval. The horizontal line separates mood categories whose ICC above or below 0.5

| Mood | Baseline | RF | | |
|---|---|---|---|---|
| | | Features | Accuracy | AUC |
| Overall | 50.8 | Verb + Facial | 69.0 | 0.75 |
| Happy | 51.9 | Verb + Facial | 64.0 | 0.70 |
| Excited | 50.8 | AVF | 68.3 | 0.74 |
| Angry | 54.9 | All | 64.4 | 0.69 |
| Disappointed | 52.3 | Verb + Visual | 66.0 | 0.70 |
| Sad | 50.8 | Verb + Audio | 64.2 | 0.62 |
| Relaxed | 54.2 | - | - | - |
| Bored | 54.2 | AVF | 64.1 | 0.65 |
| Stressed | 58.3 | - | - | - |
| Surprised | 51.5 | - | - | - |
| Nervous | 56.8 | - | - | - |

As an additional performance measure, we also present the computed Receiver operating characteristics (ROC), area under the curve (AUC). The AUC in binary classification, is equivalent to: “the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance” [Fawcett 2006]. This means that the greater the area, the better the performance of the classifier. In Table V, we observe that the Overall Mood label also results in highest AUC (0.75), and that the most reliable moods seem to obtain only slightly higher AUC (0.69-0.75) compared to the rest (0.65-0.7).

Figure 5 shows the ROC curves from the AUC values, the ROC curves are computed merging the 10 folds. We can observe that RF is a promising classifier for Happy with AUC=0.70 (confidence interval (c.i.)=[0.63,0.76]), Excited (with AUC=0.74, c.i.=[0.68,0.80]), Disappointed (AUC=0.70, c.i.=[0.63,0.76]) and Angry (AUC=0.69, c.i.=[0.63,0.75]). Moreover, the highest AUC values also correspond to the highest positively correlated moods (Happy and Excited) in Section 5. Similarly, the AUC values for Disappointed and Angry are in concordance with their observed strong positive correlations.

We conclude this section by discussing our findings in comparison with previous work:

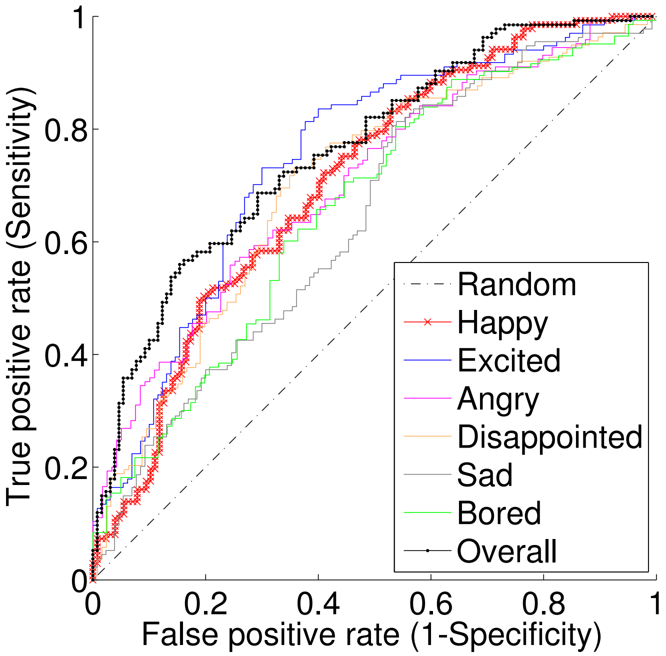


Fig. 5. AUC from best performed moods using RF (Graph best viewed in color).

- Although no direct comparison with text blogs is possible we can point out as an example the best overall performance (63.5%) obtained using word sentiment orientation, verbs, adjectives, BoW, and text statistics in [Keshtkar and Inkpen 2009].
- With respect to the acoustic channel, not direct comparison can be done, nevertheless as shown in Table I, the work in [Lee and Narayanan 2005] reported error rate of 14.1 for male speakers and 13.8 for female speakers on a word level binary classification task using several feature selection approaches. In our case, we infer the mood of the vlog (3 minutes in average) using word categories extracted from manual transcriptions and audio features from the full vlog.
- With respect to video, no direct comparison to previous work is possible, nevertheless as shown in Table I, the work in [Wollmer et al. 2013] reported performance of up to 73% (F1-weighted measure) using multimodal features to discriminate between negative and positive movie reviews. In our case, for the overall mood we obtain 69% accuracy on a binary task using more diverse data.

### 7.6. Limitations

It is important to remark that the best performing feature combinations often included verbal content features extracted from manual transcriptions. As we discussed in Section 7.3 the performance using only automatic speech transcriptions is not high enough, and thus it does not improve the performance when combine them with the nonverbal cues. This results confirm trends observed in previous works that show that automatic speech recognition (ASR) on YouTube data is still challenging [Biel et al. 2013; Hinton et al. 2012].

We must emphasize that for YouTube vlogs the mood classification task is hard for multiple reasons. First, people talk most of the time (roughly 70% of time). This strongly affects the reliability of CERT features as the face moves due to speech production, not facial expression production. Second, behavior is real and can be subtle.

Third, we address person-independent mood classification/ranking tasks, by definition more challenging (one sample per individual) than cases where multiple samples per person were available.

## 8. MOOD RANKING RESULTS

As described in Section 6.4, for the evaluation of the ranking algorithm, we estimated average precision, recall, and F1. Table VI summarizes the results for these measures (at top 10), and also for the $\tau$ rank correlation. The mood categories are organized in the same order as discussed in Section 4. For $\tau$, the results that are statistically significant ($p < 0.05$) are marked with $*$.

Table VI. Average Precision, Recall, and F1 at top 10, and Kendall correlation coefficient (N=264, *: $p < 0.05$, one-sample t-test). The standard deviation (sd) across folds is also reported. Moods are ordered according to ICC reliability.

| | AP (sd) | Recall(sd) | F1 | $\tau$ (sd) |
|---|---|---|---|---|
| Overall | 0.65 (0.23) | 0.53 (0.16) | 0.58 | 0.23 (0.18)* |
| Happy | 0.70 (0.18) | 0.58 (0.11) | 0.63 | 0.28 (0.13)* |
| Excited | 0.76 (0.19) | 0.59 (0.10) | 0.67 | 0.31 (0.14)* |
| Angry | 0.62 (0.18) | 0.55 (0.11) | 0.58 | 0.20 (0.16)* |
| Disappointed | 0.63 (0.23) | 0.49 (0.10) | 0.55 | 0.15 (0.13)* |
| Sad | 0.71 (0.17) | 0.53 (0.09) | 0.61 | 0.20 (0.07)* |
| Relaxed | 0.69 (0.22) | 0.54 (0.16) | 0.60 | 0.24 (0.14)* |
| Bored | 0.58 (0.18) | 0.46 (0.10) | 0.51 | 0.10 (0.13)* |
| Stressed | 0.69 (0.14) | 0.56 (0.11) | 0.62 | 0.21 (0.15)* |
| Surprised | 0.58 (0.23) | 0.43 (0.16) | 0.49 | 0.07 (0.16) |
| Nervous | 0.56 (0.20) | 0.43 (0.11) | 0.49 | 0.04 (0.15) |

As we can observe, the recall performance for Happy, Excited and Angry indicates that on average, about 6 vlogs are correctly retrieved in the top 10 list, and their AP indicates how early they appear in the top positions. For the cases of Overall, Disappointed, Sad, Relaxed and Bored, about 5 vlogs are correctly retrieved in the top 10 list. Finally, for Surprised and Nervous, only 4 vlogs are recovered in the top 10 list, which is not surprising considering that even for external annotators it is more difficult to score vlogs with these moods. From Table VI we can also observe that Excited has the highest F1 value (0.67), followed by Happy (0.63), Stressed (0.62) and Sad (0.61). We can also observe F1 performance between 0.55 and 0.6 for Disappointed, Angry, Overall and Relaxed mood. Finally, for Surprised and Nervous, their F1 performance is below 0.50.

Regarding rank correlation, we can observe positive correlation coefficient for Overall mood, Happy, Excited and Relaxed (0.23, 0.28, 0.31 and 0.24 respectively), which indicates that highly scored videos tend to be ranked on top positions. For Angry, Sad and Stressed, there is a moderate correlation (0.20, 0.20 and 0.21). For Surprised and Nervous, we can observe that the rank correlation is not statistically significant ($p > 0.05$).

After manually inspecting the top vlogs per mood, we observed that the ranking algorithm tends to retrieve vlogs that capture the mood correctly, and few instances of vlogs that do not correspond to the specific mood are ranked in the bottom positions of the top 10 list. Moreover, it is worth to mention that the differences among mood scores in the top 10 list, as per the annotators, are in some cases small (in the order of 0.1-0.2). Such small difference can be missed by a ranking algorithm (providing an reversed order for example), and also penalized by the correlation coefficient $\tau$. Taking into account the difficulty of the task and the inherent variability in annotator preferences, we consider that ranking is a promising approach to recover top mood vlogs for further applications.

## 9. STUDYING CORRELATION AMONG CLASSIFICATION AND RANKING METHODS

This section aims to investigate if the mood categories that are more accurate to classify are also the moods performing high when using a ranking algorithm. To explore this question, we use ranking correlation (described in the previous section) as a measure that reflects the similarity in assessing high mood performance among classification and ranking algorithms.

We estimated the rank correlation value using as reference the ICC rankings (ground truth) and the rankings of each method. In other words, based on the ICC, we ranked the 11 moods and compared this list with the ranking method (ordered from higher to lower performance), and similarly for the rest of the methods. The first rows of Table VII presents the correlation results. As we can observe, the highest ranking correlation with the ICC is binary classification with RF (0.60), which indicates that RF might have captured the inherent difficulty across moods in a more similar way as the observers.

Table VII. Kendall rank correlation values among methods, (N=11, $* : p < 0.05$).

| | Baseline | RF | AUC RF | AP@10 | Recall@10 | F1@10 | $\tau$ Kendall (allfeatures) |
|---|---|---|---|---|---|---|---|
| ICC | -0.40 | 0.60$^*$ | -0.26 | 0.46 | 0.50 | 0.46 | 0.59 |
| Baseline | | -0.32 | 0.31 | -0.40 | -0.08 | -0.21 | -0.27 |
| RF | | | -0.18 | 0.32 | 0.28 | 0.24 | 0.26 |
| AUCRF | | | | -0.08 | 0.00 | -0.04 | -0.02 |
| AP@10 | | | | | 0.70$^*$ | 0.85$^*$ | 0.75$^*$ |
| Recall@10 | | | | | | 0.87$^*$ | 0.82$^*$ |
| F1@10 | | | | | | | 0.82$^*$ |

We also computed rank correlation among classification and ranking methods. For RF and RF AUC we did not find statistically significant correlations. Moods that are more accurate to estimate for one classifier does not necessarily correspond to the most accurate moods estimated by other methods.

For AP, high and significant correlations can be observed with recall and F1 at top 10, which is not surprising since those are strongly related measures. Moreover, AP, Recall and F1 from the ranking algorithm are congruent with the ranking correlation $\tau$ as we can observe from Table VI, and confirmed with statistically significant rank correlations $\geq 0.75$. In other words, the moods Excited, Happy, Overall and Relaxed moods reflect high performance for AP, Recall and F1, as well as for $\tau$ Kendall.

## 10. POTENTIAL FUTURE APPLICATIONS

As discussed in the introduction, new generations use conversational social video for entertainment, both producing and watching content. Speculating about the future a first potential application of our work could be “mood-based recommendation lists suggestion list on YouTube. The list would enrich current discovery options, complementing existing YouTube options where vloggers are listed on the site based on their number of subscribers. This mood-based ranking could provide affective contextual information to potential subscribers, and increase the options to find personalized entertainment, e.g. allowing viewers to identify their mood with that of a particular vlogger. Interactions with vloggers who share similar emotional states in specific moments or situations, could also contribute to strengthening the vlogging communities. Finally, while a dedicated Comic YouTube channel exists in which funny vloggers participate, there are users who could benefit of ways of sharing or finding videos conveying other affective states.

A second potential application is centered on supporting video production. We anticipate two use cases. In the first one, the result of the mood impression analysis could be delivered back to vloggers to help them reflect about the perceptions they might elicit on their audience. This could support users on making decisions about what to post. The second use case is about enriching video posts. A vlog mood tracker could learn the mood variations of a user, detect mood peaks, and make suggestions to introduce sound, animations, or effects at specific moments. This could facilitate the production of certain types of vlogs, or even turn it into a fun feature for some users (and audiences.)

A third application is large-scale analysis of mood trends. As done today with tweets [Feldman 2013; Golder and Macy 2011; Mislove et al. 2010], trends of affective states could be extracted and aggregated from video to capture audience responses to political events, elections, or natural disasters. These real-time trends could then be broadcasted on dedicated channels. Video mood real-time trends on local or global matters could allow viewers to select and watch vlogs according to these mood trends, and join conversations to share their own viewpoints.

## 11. CONCLUSIONS

We presented a systematic study of mood inferences (classification and ranking) on conversational social video from verbal and nonverbal cues. Our study was based on a YouTube vlog data set that is diverse in terms of topics and people. While classification is a standard way of validating our framework, ranking is a task that in practice can have a wide applicability. We showed that while the mood classification task is challenging, several of the moods can be recognized with performance that is statistically better than a majority class baseline. The best performance was obtained for Overall mood and Excited (69% and 68% accuracy), which are categories that can be of great value in social video applications.

Our study showed that although multimodal features perform better than single channel features, not always all the available channels are needed to accurately discriminate mood in videos. We observed that the verbal content augmented the nonverbal information for many of the moods. Our work revealed that to discriminate mood it is important to know the spoken categories appearing in a vlog, including categories related to *health*, *swearing words*, *anger*, etc., in addition to the *positive* and *negative* emotion categories, that have shown improvement in mood inference from text blogs.

Several future directions can be taken. Our research has taken a system integration approach, where we have relied on existing modules (like CERT and LIWC) to conduct the study. Clearly, the integration of other recent algorithms for feature extraction could result in improved performance. Another direction includes the analysis of multiple moods per vlog. For instance, the outputs from the classifiers could feed a single model to jointly infer the various moods appearing in a vlog. In addition, the verbal content analysis could be performed using other options like WordNet Affect, instead of LIWC. Finally, individual ranking preferences could be studied, i.e., learn models based on personalized ranking (by specific annotators or audiences).

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Joan-Isaac Biel from Idiap Research Institute, and Dr Junji Yamato from NTT Comunication Science Laboratories for their contributions on early stages of this research. We also thank Dr. Trinh-Minh-Tri Do (Idiap) for discussions. This work was done in the context of the NISHA project (NTT-Idiap Social Behavior Analysis Initiative) and the SNSF UBImpressed project (Ubiquitous First Impressions and Ubiquitous Awareness).

## REFERENCES

Nalini Ambady and Robert Rosenthal. 1992. Thin Slices of Expressive Behavior as

Predictors of Interpersonal Consequences. A Meta-Analysis. *Psychological Bulletin* 111 (1992), 256–274.

Joan-Isaac Biel and Daniel Gatica-Perez. 2011. VlogSense: Conversational Behavior and Social Attention in YouTube. *ACM Transactions on Multimedia Computing, Communications* 7, 1 (2011), 33:1–33:21.

Joan-Isaac Biel and Daniel Gatica-Perez. 2012. The Good, the Bad, and the Angry: Analyzing Crowdsourced Impressions of Vloggers. In *Proceedings of International Conference on Weblogs and Social Media*.

Joan-Isaac Biel and Daniel Gatica-Perez. 2013. The YouTube Lens: Crowdsourced Personality Impressions and Audiovisual Analysis of Vlogs. *IEEE Transactions on Multimedia* 15, 1 (2013), 41–55.

Joan-Isaac Biel, Daniel Gatica-Perez, John Dines, and Vagia Tsminiaki. 2013. Hi YouTube! Personality Impressions and Verbal Content in Social Video. In *International Conference on Multimodal Interfaces (ICMI)*.

Joan-Isaac Biel, Lucia Teijeiro-Mosquera, and Daniel Gatica-Perez. 2012. FaceTube: Predicting personality from facial expressions of emotion in online conversational video. In *International Conference on Multimodal Interfaces (ICMI)*.

Paul Boersma. 2002. Praat, a system for doing phonetics by computer. *Glot international* 5, 9/10 (2002), 341–345.

Gary Bradski and Adrian Kaehler. 2008. *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.".

Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

Munmun De Choudhury, Scott Counts, and Michael Gamon. 2012. Not all moods are created equal! exploring human emotional states in social media. In *AAAI International Conference on Weblogs and Social Media*.

Oxford Dictionaries. 2014. Oxford online dictionary. (2014). http://oxforddictionaries.com/definition/english/mood

Trinh-Minh-Tri Do and Thierry Artières. 2012. Regularized bundle methods for convex and non-convex risks. *J. Machine Learning Research* 13, 1 (2012), 3539–3583.

Paul Ekman and Wallace V Friesen. 2003. *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk.

Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Commun. ACM* 56, 4 (2013), 82–89.

Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *The Journal of machine learning research* 4 (2003), 933–969.

Felix Gillete. 2014. Hollywood's Big-Money YouTube Hit Factory. Bloomberg Businness Week. (2014). Aug 28.

Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, Seyedmohammad Mavadati, and Dean P Rosenwald. 2013. Social Risk and Depression: Evidence from Manual and Automatic Facial Expression Analysis. In *Automatic Face and Gesture Recognition (FG), IEEE International Conference and Workshops on*.

Scott A. Golder and Michael W. Macy. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science* 333, 6051 (2011), 1878–1881.

Thomas Hain, Lukas Burget, John Dines, Philip N Garner, Frantisek Grezl, Asmaa El Hannani, Marijn Huijbregts, Martin Karafiat, Mike Lincoln, and Vincent Wan. 2012. Transcribing meetings with the AMIDA systems. *Audio, Speech, and Language Processing, IEEE Transactions on* 20, 2 (2012), 486–498.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep

Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and others. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE* 29, 6 (2012), 82–97.

Maurice George Kendall. 1975. Rank Correlation Methods. *London, UK* (1975).

Fazel Keshtkar and Diana Inkpen. 2009. Using sentiment orientation features for mood classification in blogs. In *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (NLPKE*.

Mark Knapp and Judith Hall. 2008. *Nonverbal Communication in Human Interaction*. Wadsworth, Cengage Learning.

Gary G Koch. 1982. Intraclass correlation coefficient. *Encyclopedia of statistical sciences* (1982).

Shiro Kumano, Kazuhiro Otsuka, Dan Mikami, Masafumi Matsuda, and Junji Yamato. 2012. Understanding communicative emotions from collective external observations. In *Proceedings of Extended abstracts, ACM Conference on Human Factors in Computing Systems (CHI)*. 2201–2206.

Chul Min Lee and Shrikanth S Narayanan. 2005. Toward detecting emotions in spoken dialogs. *Speech and Audio Processing, IEEE Transactions on* 13, 2 (2005), 293–303.

Gilly Leshed and Joseph Kaye. 2006. Understanding how bloggers feel: recognizing affect in blog posts. In *Proceedings of Extended abstracts, ACM Conference on Human Factors in Computing Systems (CHI)*.

Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett. 2011. The computer expression recognition toolbox (CERT). In *Proceedings of Automatic Face and Gesture Recognition (FG), IEEE International Conference and Workshops on*.

Gwen C Littlewort, Marian Stewart Bartlett, and Kang Lee. 2007. Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. In *International Conference on Multimodal Interfaces (ICMI)*.

LIWC. 2007. LIWC Incorporation. (2007). http://www.liwc.net/index.php

Richard Lowry. 1998. *Concepts and applications of inferential statistics*. R. Lowry.

Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, Sien Chew, and Iain Matthews. 2012. Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database. *Image and Vision Computing* 30, 3 (2012), 197–205.

François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research* 30 (2007), 457–501.

Daniel McDuff, Rana el Kaliouby, David Demirdjian, and Rosalind Picard. 2013. Predicting Online Media Effectiveness Based on Smile Responses Gathered Over the Internet. In *Automatic Face and Gesture Recognition (FG), IEEE International Conference and Workshops on*.

Gary McKeown, Michel Franois Valstar, Roderick Cowie, and Maja Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. In *Proc. ICME*.

Gilad Mishne. 2005. Experiments with mood classification in blog posts. In *Proceedings of SIGIR, Workshop on Stylistic Analysis of Text for Information Access*.

Gilad Mishne and Maarten de Rijke. 2006. Capturing global mood levels using blog posts. In *AAAI Spring symposium on computational approaches to analysing weblogs*. 145–152.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2010. Pulse of the nation: US mood throughout the day inferred from twitter. http://www.ccs.neu.edu/home/amislove/twittermood/. (2010).

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *International Conference on Multimodal Interfaces (ICMI)*.

Thin Nguyen, Dinh Phung, Brett Adams, Truyen Tran, and Svetha Venkatesh. 2010. Classification and pattern discovery of mood in weblogs. *Advances in Knowledge Discovery and Data Mining* (2010), 283–290.

Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *Affective Computing, IEEE Transactions on* 2, 2 (2011), 92–105.

OMRON. 2007. OKAO Vision. (2007). http://www.omron.com

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2, 1-2 (2008), 1–135.

James Pennebaker, Martha E Francis, and Roger J Booth. 2001. *Linguistic Inquiry and Word Count: LIWC2001*. Mahwah, NJ: Erlbaum Publishers.

James Pennebaker and Laura King. 1999. Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology* 77, 6 (1999), 1296–1312.

Dairazalia Sanchez-Cortes, Joan-Isaac Biel, Shiro Kumano, Junji Yamato, Kazuhiro Otsuka, and Daniel Gatica-Perez. 2013. Inferring Mood in Ubiquitous Conversational Video. In *Conference on Mobile and Ubiquitous Multimedia (MUM)*.

Klaus R Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech communication* 40, 1 (2003), 227–256.

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication* 53, 9 (2011), 1062–1087.

Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. 2012. Avec 2012: the continuous audio/visual emotion challenge. In *International Conference on Multimodal Interfaces (ICMI)*. ACM, 449–456.

Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. 2006. Emotion recognition based on joint visual and audio cues. In *Proceedings of International Conference on Pattern Recognition*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of Conference on empirical methods in International Conference on Natural Language Processing*. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Int. Workshop on Semantic Evaluations*. Association for Computational Linguistics, 70–74.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*. ACM, 1556–1560.

Michel F Valstar, Bihan Jiang, Marc Mehu, Maja Pantic, and Klaus Scherer. 2011. The first facial expression recognition and analysis challenge. In *Proceedings of Automatic Face and Gesture Recognition (FG), IEEE International Conference and Workshops on*.

Martin Wollmer, Felix Weninger, Tobias Knaup, Bjorn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. YouTube Movie Reviews: In, Cross, and Open-domain Sentiment Analysis in an Audiovisual Context. *Intelligent Systems, IEEE* 28, 3 (2013), 46–53.

YouTube. 2014a. YouTube Channels. (2014). http://www.youtube.com/channels

YouTube. 2014b. YouTube Statistics. (2014). http://www.youtube.com/yt/press/statistics.html

Received February 2014; revised March 2014; accepted June 2014