Mining Crowdsourced First Impressions in Online Social Video

Joan-Isaac Biel and Daniel Gatica-Perez, *Member, IEEE*

***Abstract*—While multimedia and social computing research have used crowdsourcing techniques to annotate objects, actions, and scenes in social video sites like YouTube, little work has addressed the crowdsourcing of personal and social traits in online social video or social media content in general. In this paper, we address the problems of (1) crowdsourcing the annotation of first impressions of video bloggers (vloggers) personal and social traits in conversational YouTube videos, and (2) mining the impressions with the goal of modeling the interplay of different vlogger facets. First, we design a human annotation task to crowdsource impressions of vloggers that extends a tradition of studies of personality impressions with the addition of attractiveness and mood impressions. Second, we propose a probabilistic framework using Topic Models to discover prototypical impressions that are data driven, and that combine multiple facets of vloggers. Finally, we address the task of automatically predicting topic impressions using nonverbal and verbal content extracted from videos and comments. Our study of 442 YouTube vlogs and 2,210 annotations collected in Mechanical Turk supports recent literature showing the feasibility to crowdsource interpersonal human impression with comparable quality to what is reported in social psychology research, and provides insights on the interplay among human first impressions. We also show that topic models are useful to discover meaningful prototypical impressions that can be validated by humans, and that different topics can be predicted using different sources of information from vloggers' nonverbal and verbal content, as well as comments from the audience.**

***Keywords*—Crowdsourcing, vlogs, personality, mood, attractiveness, Topic Models, nonverbal behavior, verbal content, automatic prediction**

I. Introduction

The race between Facebook and Twitter to incorporate online video in early 2013, shows an increasing interest to explore online video as a form of interaction in social media. While these type of video practices expand and diversify, there is both a need and an opportunity to study other established forms of online social video contextualizing with other types of social media. In this article, we focus on the study of conversational videoblogs (a.k.a vlogs), a form of online social video that is used by people to broadcast themselves in front of a camera, and that is one of the most popular formats in YouTube. Vlogging is a unique medium for online interaction, as it enables vloggers to produce and share a myriad of spontaneous audiovisual nonverbal cues (through face, body, and voice), that enrich the verbal content that is already found in other social media.

We address the problem of annotating vlogs with respect to vloggers' interpersonal impressions. The study of first impressions in vlogging, and social video in general, is important because viewers' behavior when sharing, commenting or rating videos is based on their impressions and reactions when watching content. In addition, this type of annotations can be used to characterize users' traits, thus opening new opportunities for user profiling, and multimedia discovery and search tasks through supervised and unsupervised automatic techniques. In this context, human annotations can be used as ground truth, because assessing human traits is a human perceptual task in nature, and because the use of short amounts of behavioral data or "thin-slices" has been documented as suitable for the study of personal and social constructs [2].

In recent research, we investigated the crowdsourcing of interpersonal and affective impressions from online video with the study of personality impressions (i.e. how people see others' personality) [8]. The advantages of using crowdsourcing in this context are twofold. First, crowdsourcing is potentially a fast and affordable method to scale human annotation to the large amount of social video available online, under the assumption that the annotation outcome is reliable. Second, by using crowdsourcing we have access to a large and diverse pool of annotators that we would not have otherwise in a traditional annotation task, as shown by the the demographic variety of annotators in crowdsourcing sites like Mechanical Turk [5].

In this article, we investigate the use of crowdsourcing to collect multifaceted impressions of a dataset of vloggers and to mine emerging patterns of these impressions along multiple dimensions. The annotation of multifaceted impressions is an opportunity to go beyond the focus of most social media literature that has investigated users' traits and states individually [20], [17], and to examine the interplay among different facets of people documented in social psychology [16]. In this regard, we argue that the three facets studied (personality, attractiveness, and mood), though not exhaustively, cover a broad range of impressions that can be built on the basis of vloggers' behavior. In addition, using a broad list of impressions we address a more realistic scenario in which people make a variety of impressions while watching online videos. This enables the discovery of impressions that are relevant to the conversational vlogging setting, are data driven, and are not limited to a small number of labels as typically done with existing social inference tasks, thus opening the door to the development of multifaceted inference models.

The contributions of this work are the following:

- We design an original task to collect crowdsourced impressions of personality, attractiveness, and mood for 442 YouTube vlogs. Our analysis of annotations revisits several key results reported in social psychology literature (with experiments mostly done in lab settings) regarding the agreement of judgments and their interplay. A result of our analysis is that social media elicits the

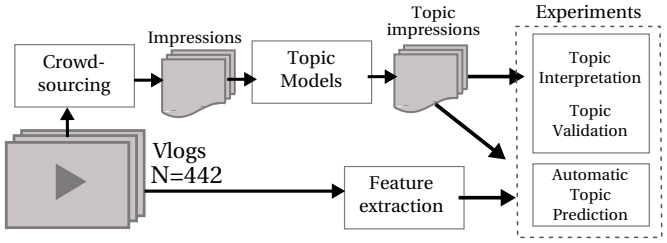

Fig. 1. A summary of our approach to study multifaceted impressions of vloggers. We use MTurk to crowdsource a diverse set of first impressions and propose the use of Topic Models to discover multifaceted topic impression by treating vloggers as documents and impressions as words. Our experiments are targeted to interpret and validate the models, as well as to address the task of topic impression prediction using automatic behavioral analysis.

same type of first impressions produced in face-to-face interactions.

- We propose a probabilistic framework to represent vloggers in the multidimensional space of impressions, using a bag-of-impressions representation and probabilistic topic models. We show that a standard Latent Dirichlet Allocation model applied on the bag-of-impressions automatically identifies meaningful prototypical impressions that are data-driven and emerge from online video watching. These prototype impressions, to our knowledge, have not been discussed previously in social media, and provide deeper understanding about the traits and states of online video producers.
- We propose a new computational task, namely automatic prediction of "topic impressions" using nonverbal and verbal content analysis extracted from vlogs. We investigate the use of several feature sets used in the literature: audio, visual, and multimodal activity cues as well as facial expressions cues (for nonverbal), and linguistic categories from transcripts (for verbal). In addition, we introduce YouTube comments as a new data source that contains information about vloggers.

We presented some preliminary results of this work in a short paper [6], where we addressed the first contribution listed above. In the current manuscript, we focus on the two other contributions, showing the relevance of this new multidimensional characterization of vloggers for both social psychology and multimedia applications.

The rest of the paper is organized as follows. Figure 1 summarizes our approach. In Section II, we review related work. Section III presents our dataset. In Section IV, we describe the process to crowdsourcing and mining annotations with topic models. Section V explains automatic feature extraction methods. We present our experiments and results in Section VI, and we conclude in Section VII.

II. Related Work

We first review the literature studying social media user traits. Second, we discuss research exploring crowdsourcing for multimodal corpora annotation. Finally, we review works addressing the characterization of human traits based on automatic behavioral analysis.

A. Characterizing Social Media Users

Our work relates to a growing body of research that has investigated several facets of social media user personality, attractiveness, and mood within interpersonal perception research and data mining.

Works investigating interpersonal perception in social media have mainly focused on the study of personality traits as broad descriptors of users' characteristics. This approach has been partly motivated by the wide acceptance of the Big-Five model: a framework to organize human personality in terms of five basic dimensions [31], the existence of standard measures of personality, as well as a tradition of personality research in social psychology. This research has measured self-reported personality and personality impressions to investigate how people present themselves in social media [20], how they convey personal information when creating their user profiles [44], and how they are seen by others based on their online behavior [20]. These works have also investigated user profiles to identify what elements from text and pictures are associated to more or less accurate personality impressions [17]. Research has also used automatic verbal analysis techniques to mine verbal content from blogs and to study the links between word usage and bloggers' self-reporter personality [50] and personality impressions [29], and has address the task if automatically predict blogger personality from text [37]. Overall, these works have backed up a general result from social psychology showing that the nature and the strength of valid personal impressions (i.e. impressions matching self-reports) varies a lot depending in the context in which impressions are made [20], [11].

Prior interpersonal perception research has also investigated attractiveness impressions in social media. In particular, research in online dating sites has studied what multimedia elements of user profiles are associated with different facets of attractiveness [19] and what user attributes (e.g. income, education, physique, religion, political inclination, etc) are associated to mate preferences [23]. The question has also been studied in Facebook, where works have analyzed the attractiveness impression information conveyed by user profiles in relation to the user activity and the attractiveness of friends [47]. Other research has shown that people prefer attractive users when initiating online relationships with zero history [48].

The large amount and the spontaneity of mood expression in social media has also generated an interest to use this data as coarse social sensors of mood at community and society scales. In blogging, works have investigated mood expression based on self-reported mood labels obtained from the bloggers themselves [34], and have also proposed approaches to estimate blogger mood using automate text analysis [36]. Other related works include mining mood expression to daily activities and social interactions [33], identifying seasonal trends on mood [3]; or exploiting the aggregates of Facebook status updates to obtain a gross happiness measure [26].

One main difference between the previous literature and the work presented here is the we focus on the video format. Yet more relevant is that we study multiple facets of users together, instead of approaching one single facet alone as most previous

research; including our previous research work personality impressions in vlogging [8].

B. Crowdsourcing Human Impressions and Social Media Annotation

Among the numerous works that have investigated the use of crowdsourcing to label multimedia corpora our work relates to research annotating human judgments from audio and video with the goal to semantically organize and retrieve data. Examples of this are the use of crowdsourcing to annotate the emotional level of speech [41], or the collection of self-reported boredom caused from watching videos [42]. In the context of social media, other research has crowdsourced the annotation required to train a machine learning system to score and track news feeds' sentiments [12], and to identify microblogs' sentiment polarity [15],

Most works on crowdsourcing personality data, have explored the possibility to administer personality questionnaires to MTurk workers for social psychology studies [13], [4]. Though the actual questionnaires and data sizes differ, these works coincide in that the quality of data met the psychometric standards associated with published research. In addition, while these samples are not representative of the broad population, they are as diverse and more representative of traditional social psychology samples [13], [5]. The work in [45] using Facebook to collect self-reported personality data from millions of users can be seen as another crowdsourcing experiment of personality data.

Our work also relates to research that has use topic models to characterize the annotation of social media content, albeit not explicitly addressing the annotation of human impressions. This research has mainly focused on modeling the large noisy vocabulary of tags available in bookmarking applications [24], [27] or discovering the semantic classes of music tags obtained through games with purpose [49], with the goal to organize social media collections and aid content recommendation and retrieval.

C. Modeling Human Behavior from Audio and Video

Our work relates to previous research studying personal and social constructs from nonverbal and verbal behavior. Regarding nonverbal behavior, several works have studied the automatic prediction of self-reported personality and personality impressions [30], [35], [28]. For example, research has investigated the classification, regression and ranking of both self-reported personality and personality impressions using audio from daily interactions [30]. In the regression tasks, best results were achieved for Emotional Stability and Extraversion impressions using prosodic cues, but could not predict self-reported personality. From audio only, the prediction of personality impressions has also been investigated in professional radio broadcasts [35], and has found that Extraversion and Conscientiousness are the highest classification performance. In face-to-face meetings [28], research has used audio cues (speech activity and prosody) and visual cues (energy from head, hands, and body) to predict self-reported extraversion and locus of control.

Regarding verbal behavior, some works have investigated language usage in transcriptions of recorded daily interactions to study the links between everyday expression and both self-reported and impressions of personality [32]. In addition, the work in [30] also investigated the fusion of text and acoustic features for the task of automatically predicting personality, and obtained better results than using prosodic features alone for Extraversion, Emotional Stability, and Conscientiousness.

Summing up, related research in social media has been prolific to understand user profiles, photos, and text as sources of personality and attractiveness information and mood expression, while works in crowdsourcing have shown the feasibility to enrich all this data with human annotations. In previous work, we have integrated these approaches for the study of interpersonal perception in vlogging. In particular, we investigated the feasibility of crowdsourcing personality impressions and addressed the task of predicting Big-Five personality impressions by focusing on the nonverbal [8], [9] and verbal behavior of vloggers [7], also contributing to research on human behavior from audio and video. In the current manuscript, we address the study of vlogger impressions by proposing a method to mine impressions that are multifaceted. This is novel with respect previous research because by studying impressions other than personality we expand first impressions research to other user traits that may be equally important in vlogging, and at the same time, we approach a more realistic scenario in which human impressions are multidimensional. Moreover, the discovery of these new multidimensional user categories is an important problem in social psychology research interested in identifying basic factors describing human traits, as it is exemplified by the large body of literature focused on discovering the Big-Five personality traits. Finally, the use of unsupervised methods to discover user categories is also relevant for new multimedia applications.

III. YouTube Vlog Dataset

We use a dataset consisting of 442 YouTube vlogs that was previously used in [8]. All these videos show one single person in front of a webcam talking about diverse topics including personal issues, politics, movies, books, etc. Apart from this, the collection was not restricted to any type of content, so the language and the behavior in the videos are natural and diverse. The collection is mostly balanced in gender, with 208 males (47%) and 234 females (53%). To reduce the time required by annotators to watch vlogs, the videos are shorten to the first conversational minute as described in [8]. The use of short amounts of interaction or "thin-slices" has already been documented in psychology as suitable for the study of first impressions [2], [39].

IV. Mining crowdsourced vlogger impressions

We first present the design of the crowdsourcing task. Second, we introduce our approach to use Topic Models to discover multifaceted impressions.

A. Crowdsourcing Task

We crowdsourced the annotation of vlogger impressions using Amazon's Mechanical Turk. The task was designed to resemble the video watching experience in YouTube, where people are exposed to large amounts of content, and based on the impressions made from playing bits of a video, decide whether is worth watching or not.

Our Human Intelligence Task (HIT) consisted of two main components. We designed an embedded video player to display the one-minute vlog slices obtained from preprocessing (see Section III). The bottom part of the HIT included four questionnaires used to assess personality, attractiveness, mood, and demographics of vloggers. With the purpose of obtaining spontaneous first impressions, we did not give any particular instructions to workers to fill the questionnaires apart from (1) watching the video entirely, and (2) answering the questionnaires. Table I summarizes the traits annotated using the four questionnaires, which we describe as follows.

1) Personality questionnaire: We annotated the personality impressions of vloggers using the Ten-Item Personality Inventory (TIPI) designed by [21]. The TIPI measures the Big-Five traits of personality by means of 10 items (two items per scale on a 7-point likert scale), and is an instrument specially thought to be used when time is limited (it can be completed in approximately one minute). The questionnaire has shown reasonable psychometrics with respect to longer personality tests, and has already been used in several works to measure personality in social media settings [20], [43]. In our case, we decided to use the TIPI in order to keep the annotation task as brief as possible. As suggested by [20], The TIPI instructions were to ask workers about the vlogger personality impressions. The Big-Five are summarized in Table I and the TIPI questionnaire is available in the supplementary material.

2) Attractiveness questionnaire: We did not find any standard, short instrument in the literature to report attractiveness impressions, and therefore we designed our own. Our questionnaire was inspired by research investigating attractiveness from physical and non-physical facets [19], [25]. First, we gathered a list of five facets that cover different aspects of attractiveness judgments: two facets of physical attractiveness (beautiful and sexy), and three facets of non-physical attractiveness (likable, friendly, and smart). Then, we phrased five items similarly to the personality questionnaire, using two adjectives to describe each facet, and a 7 point likert-scale. Finally, we added a sixth item to annotate the overall attractiveness of the vlogger. The six attractiveness items are summarized in Table I, whereas the attractiveness questionnaire can be found in the supplementary material.

3) Mood questionnaire: We designed our own mood impression questionnaire based on existing literature investigating mood in blogs [34]. The process consisted on iteratively shorten a list of 132 moods found in the LiveJournal blog dataset to obtain a list of 20 mood terms that paired together would be useful to describe 10 different moods states. The idea was to keep those moods that were more likely to manifest in a conversational setting like vlogging, and thus, easier to judge by people (specially from nonverbal behavior), and that at the same time would cover a range of arousal and pleasure levels. The resulting questionnaire included three positive moods (from high to low arousal): excited, happy, relaxed; two neutral moods: bored, surprised; and five negative moods (from high to low arousal): anger, stressed, nervous, sad, and disappointed. Finally, we added an eleventh item to annotate the overall mood of the vlogger on the pleasure dimension (i.e. negative/positive mood). All moods were annotated using a 7 point likert-scale. The eleven mood items are summarized in Table I, whereas the mood questionnaire can be found in the supplementary material.

Questionnaire	Trait
Personality	Big-Five: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience
Attractiveness	Beautiful, Likable, Friendly, Smart, Sexy, Overall attractiveness
Mood	Happy, Excited, Relaxed, Sad, Bored, Disappointed, Surprised, Nervous, Stressed, Angry, Overall mood

TABLE I. SUMMARY OF THE CROWDSOURCED ANNOTATIONS.

The actual design of the HIT resulted from an iterative process to make sure that instructions were clear; HIT payment was competitive with existent MTurk tasks; and most importantly, to discourage spammers by making sure that workers watched the entire video before they could answer the questions. In addition, we restricted HITs to workers from US and India, to make sure they speak English; and with HIT acceptance rates of 95% or higher, to ensure a minimum qualification from workers,

Running the MTurk task required substantial control, answering emails from workers, ensuring that they understood the task, and accepting submitted HITs in order to built a community of engaged and trusted workers. In total, we posted 2,210 HITs paid at $\$0.15$ to annotate five times each one of the 442 vloggers. The whole task was completed after 14 days of submission and the the average time spent by workers on the HIT was 256 seconds (and a minimum of 89 seconds). Given the timings obtained, the control mechanisms of the HIT design, and the burden of further validating annotations given the perceptual nature of the task, we did not enforced any filtering, and rather aggregate the annotations across the five MTurk workers for each vlogger. Details about the demographics of the workers are not reported here for space reasons but can be found in [6].

B. Mining Multifaceted Impressions With Topic Models

We propose the use of the Latent Dirichlet Allocation (LDA) model to discover joint first impressions of vlogger personality, attractiveness, and mood. We used LDA because its usefulness on characterizing documents rather than for its ability to cluster data (which could be otherwise achieved with other techniques). In addition, the probabilist framework of LDA is useful to develop multimedia applications such as retrieval or comparison, and is highly modular, i.e. allows future model improvements, for example to characterize impressions together with annotators.

LDA is a probabilistic generative model originally designed for discovering topical patterns in text documents based on word co-occurences [10]. The LDA model is also represented

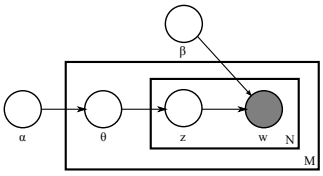

Fig. 2. Graphical model representation of LDA. The boxes are plates representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

as a probabilistic graphical model in Figure 2. A document is generated by first sampling a distribution over topics θ_d. Then, for each word, a topic z is drawn and a word w_n is sampled from $p(w_n|z_n, \beta)$, i.e., the distribution of words conditioned to topic z. The distributions of documents over topics θ_d and the distribution of words β are learned from the data [10]. The basic idea in LDA is that documents are represented as random mixtures over latent topics, and topics are characterized by a distribution over words. This is formally represented by the marginal probability of a document w as:

$$p(w|\theta, \beta) = \sum_z p(w|z, \beta)p(z|\theta) \quad (1)$$

where z denotes topic, θ is a distribution of documents over topics, and β is the parameter of Dirichlet prior in the per-topic word distribution.

LDA helps to explore document collections, by interpreting topics based on their most likely words and visually representing them using their most likely documents, or by characterizing documents with their most likely topics.

Because LDA is designed to exploit term-frequency, we have to explicitly use some terms to represent the low-scores of personality and attractiveness, as otherwise the model would not capture these traits when given very low values. Note that this is fine, for moods, that are not always present, but it is not for personality or attractiveness impressions. To address this, we used two different words to represent each personality and attractiveness trait (e.g. low-extr and high-extr for Extraversion) and one word to represent moods. In total, our vocabulary consists of 33 unique words that are summarized in Table II. For each vlogger, the bag-of-words was built by replicating each word proportionally to the aggregate scores given to the corresponding impression, as follows. High personality, high attractiveness, and mood words counts are determined by $c_w = \sum_{i=1}^{k} s_i(w) - ks_{min}$, were w is the word, k is the number of annotators ($k = 5$), $s_i(w)$ is the score given by annotator i when judging the trait associated with word w, and s_{min} is the lowest score in the likert scale ($s_{min} = 1$). Similarly, low personality and low attractiveness word counts are determined by $c_w = ks_{max} - \sum_{i=1}^{k} s_i(w)$, where s_{max} is the maximum score in the likert scale. Overall, we obtained 442 documents and 146,738 word counts. Also note that by construction, each word appeared at least one time in every document.

V. Automatic Feature Extraction

We extracted three different sets of features to characterize vlogger behavior. The first two have been already used in previous work and model the nonverbal and verbal aspect of vlogger behavior, respectively. The third feature set characterizes vlogger behavior based on what the audience says about vloggers in their comments.

Facet	Words
Personality	high-extr, low-extr, high-agr, low-agr, high-cons, low-cons, high-emot, low-emot, high-open, low-open
Attractiveness	high-beautiful, low-beautiful, high-like, low-like, high-friendly, low-friendly, high-smart, low-smart, high-sexy, low-sexy, high-over.attr, low-over.attr
Mood	happy, excited, relaxed, sad, bored, disappointed, surprised, nervous, stressed, angry, over. mood

TABLE II. Summary of Vocubulary.

A. Nonverbal cues

We extracted a set of features that was previously used in [8] and that is composed of **audio, visual, and multimodal nonverbal cues** (thereby referred to as AV). From the audio channel, we computed 3 speaking activity features on the basis of speech-non-speech segmentations (speaking time, # speaking turns, length of speaking segments) and 98 aggregate statistics from frame-by-frame estimates of prosodic cues such as pitch, energy, and speaking rate. From the visual channel, we computed 3 looking activity cues based on looking-non-looking (looking time, # looking turns, length of looking segments), 2 pose cues (distance to the camera and vertical framing), and 5 statistical aggregates of weighted motion energy images (weighted motion energy images) that measure the accumulated amount of motion through the video. Finally, we built a multimodal segmentation based on speaking and looking segmentations and computed 3 multimodal cues: looking-while-speaking time (L&S), looking-while-not speaking time (L&NS) and the multimodal ratio (L&S/L&NS). In total, these are a total of 130 cues. A detailed description of these features can be found in [8].

B. Verbal cues

We computed verbal cues from the manual transcriptions of vlogs, performed by a professional company. As in [7], we used the Linguistic Inquiry and Word Count (LIWC) software to compute lexical features form verbal content. The LIWC is a word categorization tool that links **linguistic and paralinguistic categories** to psychological constructs and personal concerns, and has been developed and psychometrically validated in social psychology research [38]. The English version of LIWC uses a dictionary composed of 4,500 words and word stems. During transcript processing, each word is looked up in the dictionary, and in case of a match the appropriate word category is incremented (note that words can be assigned to more than one category at a time). Finally, counts are divided by the total number of words in the transcript. Since LIWC is designed to process raw text, there is no need for any type of preprocessing.

In total, we considered 65 LIWC cues, discarding 12 punctuation categories that are not relevant in the spoken setting, and we included one general descriptor that counts words longer than six letters. Table III summarizes the amount of data for manual transcriptions, including raw data (words), and LIWC

	Words	LIWC
# Terms	10K	65
# Tokens	246K	221K

TABLE III. NUMBER OF UNIQUE TERMS AND TOKENS IN MANUAL DATA: RAW VOCABULARY (WORDS) AND DATA PROCESSED USING LIWC.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Comments	1	5	14.50	85.69	54.5	999
Words (all)	2	78	256	1609	997	37441
Words (200)	2	78	256	882.3	982	11035

TABLE IV. SUMMARY STATISTICS OF COMMENT THREADS IN THE SUBSET OF 372 VLOGS WITH COMMENTS. VALUES DID NOT CHANGE W.R.T. THE FULL DATASET. (ALL) INDICATES ALL AVAILABLE COMMENTS, (200) CONSIDERS ONLY THE 200 MOST RECENT COMMENTS PER THREAD.

outputs. 91.7% of the words from manual transcripts were found in the LIWC dictionary.

C. Comments

We computed verbal content cues from YouTube comments. Though we are not aware of any work providing a general picture of the type of comments found in YouTube, we argue that comments can potentially reveal information about vloggers. For example, comments can refer to the vlogger discourse, to other previous comments, to the vlogger nonverbal behavior or looks, as well as to technical aspects from video editing or quality. In this context, work on automatic categorization of video showed the use of comments to improve the performance of a baseline classifier using audiovisual analysis and metadata [18], that supports the conjecture that comments contain useful information to characterize the video content (or the verbal content of vloggers, in our case).

Table IV summarizes some basic statistics of the set of 442 videos and their comments. Note that 70 videos did not have any comments at the time of data collection and therefore, these numbers are computed with respect to the other 372 videos. The median video had 14.50 comments and 256 words, compared to the 469 words of the median video transcription. However, the number of comments and words varies a lot across videos, as popular videos have very long comment threads. Considering only the most recent 200 comments for each video, the total text corpora from comments sums up to 328,207 words, which is 34% more data than the full transcription dataset (see Table III).

We explored two different feature sets using **LIWC** and **unigrams** (we did not use bigrams, because vectors become too sparse, which is a problem given the number of available samples). We represented each vlog with one single document including all comments (we limited long threads to the 200 most recent comments). Then, we processed comments to remove punctuation, to remove repeated letters in words (i.e. "aweeesooomeeee" was converted to "awesome"), and to stem words using Porter's algorithm. For LIWC, we used the 65 linguistic categories as features. For unigrams, we considered those that appeared in more than 10 documents, which resulted in 1286 unique words.

We observed the number of words found in the LIWC dictionary when processing comments increased from 67% to 84% after removing consecutive repeated letters. This percentage is still far from the 92% found when processing manual transcriptions with LIWC, and indicates that despite the amount of data available in comment threads, the text is very noisy and it includes many typos/misspells that result from fast, spontaneous writing.

VI. EXPERIMENTS AND RESULTS

We first examine the output of the crowdsourcing task. Second, we present and discuss the results of using topic models in our annotations. Third, we perform a topic model validation experiment using human annotations. Four, we address the task of automatic impression topic prediction. Note that any time we refer to personality traits, facets attractiveness, and moods we are not referring to the actual traits of vloggers, but to the impressions that annotators made from them.

A. Analysis of Crowdsourced Impressions

Table V shows a set of descriptive statistics about the impressions collected in MTurk: mean, standard deviation, minimum, maximum, and skewness, in addition to an intraclass correlation coefficient (ICC, a measure the level of absolute agreement between annotators computed based on ANOVA [40]).

As observed from the minimum and maximum scores, all the annotations span fully across the 7-points likert scale, which indicates that all the personality traits, attractiveness facets, and moods are found in the vlogging setting to some extent. The distribution of all personality traits and attractiveness facets are centered on the positive side of the likert scales (Mean ≥ 4) and showed little skewness (Skew $\leq \pm 1$), as it happens with positive moods (Happiness, Excitement, and Relax). In contrast, the rest of moods (negative and neutral) are centered low on the negative part of the scale and result positively skewed (≥ 1). Independently of their ICCs, it is apparent that these moods are less frequent in vlogging, or that people in these states might be less likely to make a video.

A core question of interest is to what extent workers are able to achieve any agreement on the basis of watching one minute slices of vlogs. In our setting, no agreement could result from three hypothetical situations in which either vloggers' behavior would not convey any impression information; MTurk workers did not pay attention while completing the HITs; or they simply disagree on their first impressions. We computed the Intraclass Correlation Coefficients (ICCs) for each personality trait, as they are commonly used in psychology to measure the level of absolute agreement between annotators [40]. Note that, in contrary to other existing reports of annotators agreement of personality [20], [46], we cannot use the ICC(2,k) measure, because each observer only annotated a subset of the data. Instead, we computed ICC(1,k) which is a measure of absolute agreement designed for experimental settings where each target is annotated by k judges randomly selected from a population of K judges, with $k < K$ [40]. In our setting, we have $k = 5$ and $K = 113$. The last column of Table V shows the ICC(1,5) resulting of aggregating annotations across the 5 workers. Overall, the ICC(1,5) showed moderate reliabilities for all

Trait	$Mean$	SD	Min	Max	$Skew$	ICC
Extraversion	4.61	1.00	1.90	6.60	−.32	.77
Agreeableness	4.68	.87	2.00	6.50	−.72	.65
Conscientiousness	4.48	.78	1.90	6.20	−.32	.45
Emotional Stability	4.76	.79	2.20	6.50	−.57	.42
Openness to Experience	4.66	.71	2.40	6.30	−.09	.47
Beautiful	4.41	1.02	1.40	6.80	−.48	.69
Likable	4.98	.80	2.20	7.00	−.51	.44
Friendly	5.13	.83	2.20	6.80	−.67	.51
Smart	4.74	.74	2.80	6.80	−.19	.35
Sexy	4.06	1.14	1.00	7.00	−.32	.60
Overall attractiveness	4.48	.93	1.20	6.60	−.49	.61
Happy	4.32	1.18	1.20	7.00	−.39	.76
Excited	4.54	1.20	1.20	6.80	−.39	.74
Relaxed	4.22	.93	1.60	6.20	−.50	.54
Sad	2.17	.99	1.00	6.60	1.49	.58
Bored	2.41	1.04	1.00	6.80	1.20	.52
Disappointed	2.38	1.11	1.00	6.43	1.02	.61
Surprised	2.51	.99	1.00	6.40	1.09	.48
Nervous	2.37	.82	1.00	5.20	.84	.25
Stressed	2.24	.93	1.00	6.40	1.09	.50
Angry	2.15	1.10	1.00	6.60	1.68	.67
Overall mood	4.83	1.04	1.60	7.00	−.58	.75

TABLE V. BASIC DESCRIPTIVE STATISTICS OF VLOGGER IMPRESSIONS AND INTRACLASS CORRELATION COEFFICIENTS ICC(1,K). ALL ICCS ARE SIGNIFICANT WITH $p < 10^{-3}$.

	1	2	3	4	5	6	7	8	9	10	11	12
1 Extr												
2 Agr	.04											
3 Beautiful	.20	.30										
4 Friendly	.35	.57	.54									
5 Sexy	.17	.28	.82	.50								
6 Happy	.47	.38	.37	.52	.37							
7 Excited	.64	.26	.33	.49	.33	.74						
8 Relaxed	−.12	.40	.25	.37	.28	.34	.15					
9 Sad	−.39	−.32	−.15	−.34	−.12	−.36	−.37	−.10				
10 Bored	−.40	−.30	−.18	−.35	−.14	−.26	−.39	.03	.63			
11 Disapp	−.29	−.38	−.13	−.29	−.10	−.43	−.35	−.18	.74	.51		
12 Stressed	−.28	−.34	−.14	−.30	−.11	−.31	−.27	−.20	.71	.50	.68	
13 Angry	−.11	−.58	−.15	−.35	−.12	−.35	−.20	−.25	.56	.42	.67	.60

TABLE VI. PAIR-WISE CORRELATIONS OF SELECTED IMPRESSIONS. WITH ICC(1,K) > .50 (EXCEPTUATING ABSOLUTE VALUES LOWER THAN $r = .10$, ALL CORRELATIONS ARE SIGNIFICANT WITH $p < 10^{-3}$).

personality traits (.42 < ICC(1,5) < .77), attractiveness facets (.35 < ICC(1,5) < .69), and most moods (.48 < ICC(1,5) < .76) with the exception of Stressed (ICC(1,5) = .25).

We emphasize that the magnitude of the personality impression reliabilities compares well to other personality impression works in social computing, and indicates that overall there is substantial agreement on the personality impressions from MTurk. For example, Ambady et al. [1] found that single personality impressions based on face-to-face interactions achieved reliability between .07 and .27 for different traits, whereas Gosling et al. measured reliabilities between .23 and .51 for single impressions from bedrooms using the same TIPI questionnaire [22]. Because these reliabilities were reported in terms of mean pair-wise correlations between raters, we computed these measures on our data for comparison (we considered only those annotators with more than 5 completed HITs). The resulting mean pair-wise correlations are: Extraversion (.44), Agreeableness (.36), Conscientiousness (.23), Emotional Stability (.27), and Openness to Experience (.24). Other works studying user profiles and websites reported reliabilities in terms of ICC(2,1) and may not be compared directly [46], [20].

Regarding the attractiveness and mood annotations, most of the reliabilities were comparable to those of personality traits (see Table V). Physical attractiveness facets such as Beautiful (.69) and Sexy (.61) achieved more agreement than non-physical facets like Friendly (.51) and Likable (.44). The overall attractiveness (.61) also achieved moderate agreement. Though we could not find any references in the literature to contrast these findings with, it seems clear that non-physical impressions may require more information (e.g., longer observations) than first sight judgments of physical attractiveness. Interestingly, mood impressions were on average the impressions that achieved higher agreement compared to personality and attractiveness, with the exception of Nervous (.25). High arousal moods such as Happy (.76), Excited (.74), and Angry (.67), as well as the overall mood (.75) achieved the highest agreement. This result is likely associated to the amount of visual and acoustic activity of these mood states compared to low arousal moods.

Finally, we evaluated the extent to which vlogger impressions are associated to each other by means of pair-wise correlations (Table VI). For this analysis, we focus on traits and states that showed the largest agreement amongst all (for space reasons we consider impressions with ICC(1,k) $> .50$), and also leave apart the overall attractiveness and overall mood annotations. We found a number of positive and negative effects that may be explained by a well-documented halo effect that suggests that attractive people are typically judged as holding more positive traits than unattractive people, with some exceptions [16]. For example, we found significant positive correlations between judgments of attractiveness and Extraversion (Beauty, $r = .20$, Friendliness, $r = .35$, and Sexiness $r = .17$), which have been previously reported in the literature in other settings [11]. In addition, we found that Beauty is positively correlated with positive moods (Happiness, $r = .37$, Excitement $r = .33$, Relax $r = .25$), and negatively correlated with negative moods (Sadness, $r = -.15$, Boredom, $r = -.18$, Stress $r = -.14$, and Anger $r = -.15$). This halo effect may as well be mediating some of the correlations between Extraversion and moods (Happiness, $r = .47$ or Stress, $r = -.28$). It is important to highlight that, compared to Extraversion, Agreeableness shows even stronger correlations with attractiveness and mood (e.g. Beauty $r = .30$, Friendliness, $r = .57$, Hapiness $r = .38$, Anger $r = -.58$), which to the best of our knowledge may have not been observed in the literature because Agreeableness typically achieves less agreement in scenarios different than vlogging [11]. Importantly, note that judgments of Extraversion and Agreeableness are not correlated ($r = .04, p = .30$). Finally, it is worth commenting the different associations between Relaxed and Extraversion ($r = -.12$), and between Relaxed and Agreeableness ($r = .40$), which is the only mood that shows opposite sign effects with these traits. Likely, in the first case, Relaxed was interpreted as calmed (opposite to excited), whereas in the second case it may have been judged as pleasant.

To summarize, our analysis showed that several judgments of the three dimensions are correlated to each other. However, it could also be that the links between traits are more complex than the linear associations investigated in this section. This issue would have to be studied in the future.

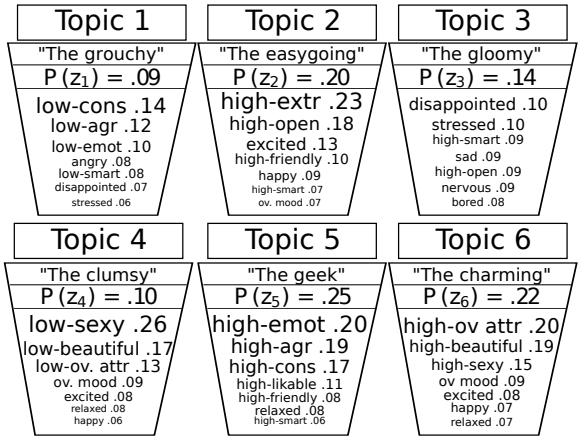

Fig. 3. Discovered LDA topics. The titles on top of each topic are suggestions of "personae" that might capture the joint meaning of the top words. Topic names are grouchy, easygoing, gloomy, clumsy, geek, charming, and were chosen by the authors using popular jargon (i.e. not formal definitions).

B. First-impression Topic Interpretation

We used Gibbs sampling to infer the distribution of documents over topics θ_d and the distribution of topics over words based on our collection of 442 vloggers. We explored different number of topics ($T = 4, 6, 8$ and 10), while setting the LDA hyperparameters to standard values ($\alpha = .1$, $\beta = \frac{50}{T}$) [10]. In general, results with very few topics have no purpose, whereas using too many topics (over 10) the model stops capturing the interplay between impressions, due to the small size vocabulary. We found that $T = 6$ gives a variate yet manageable number of topics for interpretation and discussion. Note that the results reported as follows have been replicated in smaller samples of the dataset leaving out a random 10% of the vloggers, which indicates that the topical representation output is stable (the results with left-out data are not shown here for space reasons). The labels manually given to topics in Figure 3 are strictly used as handles to facilitate the discussion, and are not formal definitions.

Figure 3 shows the top 7 words for each topic, together with their probability (the font-size of top words is proportional to the probability). It also shows the probability of each of the six topics $P(z)$. We noted that most topics resulted from a combination of personality, attractiveness, and mood impressions together, with the exception of Topic 4 and Topic 6 which are mainly characterized by attractiveness and mood impressions.

Grouchy, Easygoing, and Geek ranked multiple personality impressions among the most likely words. In particular, Grouchy is characterized by low personality impressions of Conscientiousness ("low-cons"), Agreeableness ("low-agr"), and Emotional Stability ("low-emot). It also includes low judgments of intellectual attractiveness ("low-smart"), and negative moods such as anger ("angry") or disappointment ("disappointed"). In comparison, Easygoing is characterized by high personality scores on Extraversion ("high-extr") and Openness to Experience ("high-open"). In addition, it includes high scores on nonphysical attractiveness ("high-friendly", "high-smart"), and impressions of positive moods ("excited", "happy"), and of overall mood ("ov. mood"). Compared to Grouchy, Geek is dominated by the high counterparts of the same personality traits: "high-emot", "high-agr", and "high-cons". It also included judgments of nonphysical attractiveness such "high-likable", "high-friendly" and positive mood "relaxed".

In contrast to the topics above, Gloomy represents a vlogger that is seen as disappointed, stressed, sad, and as a smart person ("high-smart"). This topic has "high-open" as the only personality judgment with high probability. Overall, this captures a combination of socially desirable high scores on personality with high scores of negative moods. Finally, Clumsy and Charming are dominated by attractiveness and mood judgments. Topic 4 is characterized by low judgments of physical attractiveness ("low-sexy", "low-beautiful","low-over.attr"), and overall positive moods ("excited", "relaxed", "happy"). Finally, Charming is characterized by high scores of physical attractiveness ("high-ov.attr","high-beautiful","high-sexy"), and positive moods such as "ov.mood", "excited", and "relaxed".

As expected, some of the correlations between impressions found in the previous subsection were captured by the LDA as word concurrencies in the topics. For example, we found that personality impressions from Conscientiousness, Agreeableness, and Emotional Stability (which were correlated with $.41 < r < .64$) had words co-occurring in both Grouchy and Charming, and did not co-occur with any other personality trait words in any other topic. In addition, Extraversion impressions, which correlated only with Openness to Experience ($r = .42$) co-occurred only with the latter in Easygoing. In the case of Charming, we found high attractiveness together with positive moods, which may result from a positive halo effect, as discussed in the previous subsection. We also found that impressions of overall attractiveness are more likely to co-occur with physical impressions of attractiveness, as seen in Clumsy and Charming, which concurs with literature showing that physical facets have larger contribution that nonphysical facets on judging attractiveness [25]. The LDA model is also useful to capture some non-linear relationships that result from the interplay between personality, attractiveness and mood . For example, we find "excited" to co-occur with "low-beautiful" (in Clumsy) and "high-beautiful" (in Charming); or the words "happy" and "sad" to co-occur both with "high-open" of Easygoing and Gloomy, respectively. Finally, we found that specific word co-occurrences may add subtle connotations to specific judgments. This occurs with impressions of disappointment, which co-occur with "angry" and "low-smart" of Grouchy and with "sad" and "high-smart" of Gloomy.

Finally, we show that vloggers are indeed modeled as mixtures of topics, rather than being represented by one single topic. To measure this, for each vlogger, we counted the number of topics that accounted for 80% of the probability mass and plot the complementary cumulative distribution in Figure 4. The plot shows that all vloggers need at least 2 topics to be characterized, and that 96% and 30% of them need at least 3 and 4 topics respectively. Figure 4 also shows two examples of a vlogger represented with 2 topics (Grouchy and Clumsy) and another represented with 6 topics, which

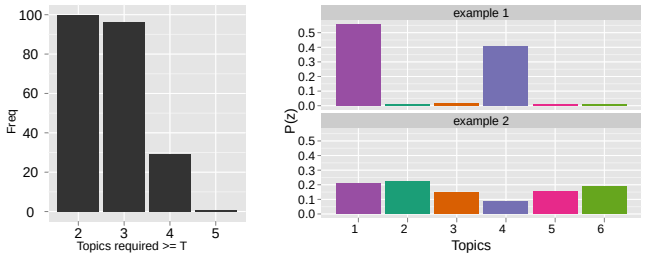

Fig. 4. Left: percentage of documents that need more than T topics to cover 80% of their topic probability mass. Right: two examples of vloggers represented as a mixture of topics. The tops case is represented well by 2 topics, and the bottom is closer to a uniform distribution.

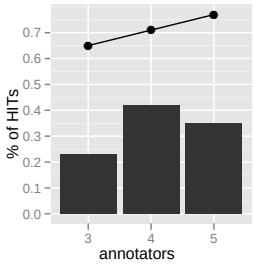

Fig. 5. Topic intrusion task summary. Bars show breakdown of agreement as measured by the number of annotators that achieved majority voting. Points show proportion of tasks where topic intruder was correctly annotated when choosing majority voting.

illustrates the diversity of the dataset.

C. First-impression Topic Validation

One fundamental question is whether the discovered topics are meaningful for a typical social video viewer. To investigate this, we validated the topic model output using the approach proposed in [14]. In this work, the authors proposed a formal setting to use human judgments to validate the two components of topic models applied to text: the composition of topics and the association between topics and documents. The first component is evaluated through a *word intrusion* task that examines to what extent a topic has human-identifiable semantic coherence, whereas the second component is evaluated through a *topic intrusion* task that examines how well the topic decomposition agrees with human judgments. While this approach is sensible for text collections (e.g news or articles), we believe that for our case of human impression topics, the task of *word intrusion* is very difficult, because there is no clear shared knowledge of what human traits are more likely to appear together, as opposed to what happens with co-ocurring words in text documents. Thus, in our experiments, we focused on the task of *topic intrusion*.

The topic intrusion task proposed in [14] consists of showing annotators a document together with a small number of associated topics (each topic is represented by the most likely words for that topic). All topics but one are the highest-probability topics assigned to that document by the topic model, while the one intruder topic is chosen randomly from the remaining low-probability topics. Then, annotators are asked to identify the intruder topic following their own judgment, and the performance of the model is measured based on the agreement between the model and human judgments.

Given the limited number of topics in our experiments (T=6) compared to the hundreds used in [14], we decided to study topic intrusion by showing only two topics per vlog: one high-probability topic, and one low-probability topic. However, to compensate for this simplification, we decided to annotate each vlog more than once using different topic combinations as follows. First, for each vlog, we identified the number of topics needed to cover up to 80% of the topic probability mass. Then, the topic intrusion task was replicated by taking each of these topics as the high-probability topic, and randomly choosing the low-probability topic from the topics covering the residual 20% of the topic probability mass. The annotation task itself consisted on watching a vlog and then deciding what of the two topics was better describing the person in the video. We do this instead of asking the intruder topic because the task becomes simpler for annotators and it is straightforward to identify the intruder by inverting the annotations.

We carried out our experiments in Mechanical Turk for a total of 36 vlogs chosen from the top documents representing each topic (we selected 3 males and 3 females for each of the 6 topics). We payed $0.03 per HIT. Vlogs were annotated between one and three times each with different topic combinations up to a total of 100 topic intrusion tasks (one task per HIT). In addition, each HIT was assigned to 5 different annotators, which sums up a total of 500 HITs. Then, we used majority voting to aggregate the annotations to obtain one topic intruder per task.

Bars in Figure 5 show the breakdown of HITs based on the agreement achieved on the topic intruder task as measured by the number of annotators that achieved the majority voting consensus (X-axis). Note that by construction the annotation task, a majority vote is always found. The figure shows that 35% of the HITs had full agreement, and that only only 23% of the HITs had the minimum agreement possible of 3 annotators. Overall, these values indicate that the majority voting on topic intruder is reliable.

The annotation agreement, however, does not tell whether the majority vote matches the topic intruder ground-truth as we constructed it. Thus, points in Figure 5 show proportion of tasks where topic intruder was correctly annotated when choosing majority voting. Note that, in the figure, these proportions are breakdown depending on the annotator agreement. In total, in 69% of the tasks (not shown in the figure) the topic intruder is correctly identified by human. Figure 5 shows that the percentage of correct tasks increases with the majority voting annotation reliability, achieving up to 77% for tasks with full reliability. These results clearly validate the quality of vlog-topic assignments found by the topic model.

D. Predicting Topic Impressions Automatically

In this section, we address the task of automatically predicting the topical impressions discovered with LDA. Our experiments aimed to evaluate the extent to which vlogger topic probabilities can be predicted automatically compared to predicting traits/states individually, and to identify what

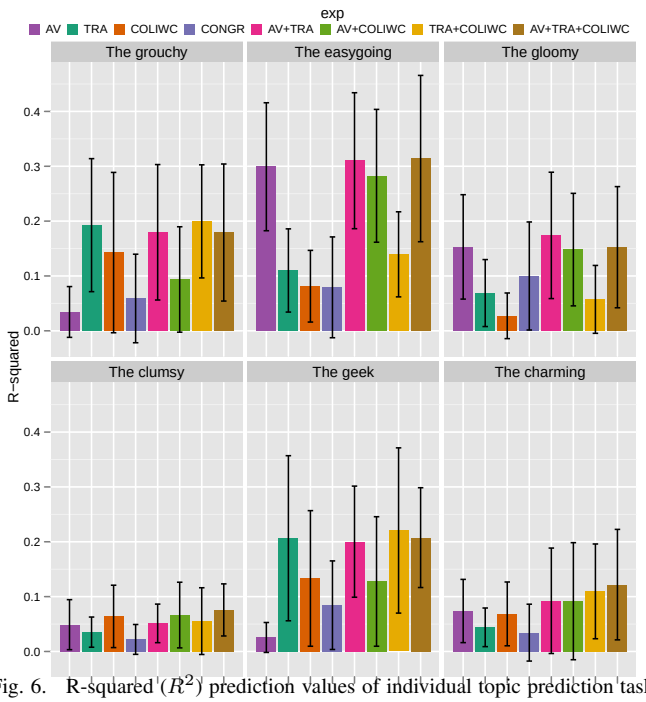

Fig. 6. R-squared (R^2) prediction values of individual topic prediction tasks using audiovisual (AV), verbal content from transcripts (TRA), comments (COLIWC and CONGR), and different combinations of them (please view in color).

behavioral channels carry useful information for this. We approached the problem using one independent regression task per topic and taking topic probabilities as target scores.

We trained and tested different models using audiovisual features (AV), vlog transcriptions processed with LIWC (TRA), YouTube comments using both LIWC (COLIWC) and unigrams (CONGR), and several combinations of them. The set of 408 vlogs was reduced to the 372 videos that have comments. In the experiments, we use a Random Forest model with features selected from a regularized Random Forest (RRF). To avoid overfitting on the selection, we use a 10-fold cross validation and run the RRF on the training data only for each fold. Results using inner selection with RRF slightly outperformed experiments using all features (i.e. no selection), so we decided to only report the former. Figure 6 shows R-squared (R^2) performance values for all topics and feature sets. The R^2 is measured based on the ratio between the models absolute error and the error obtained using a baseline that predicts the average value.

1) Audiovisual features: The highest performance of a single feature set was achieved using AV for Easygoing (R^2 = .30). With lower performance, AV was also useful to predict Gloomy (R^2 = .15), and Charming (R^2= .11). Figure 7 showing the importance of the AV features for all these topics as "tag" cloud helps with interpretation. Looking at the topic representations of Figure 3, we found that the most likely words for Topic 2, "high-extr" and "high-open", are associated with the two best predicted personality traits using AV in related literature [8]. In addition, the third most likely word in Easygoing, "excited", is a high arousal positive emotion that can potentially be measured using the wMEI

Fig. 7. RF feature importance plotted as a tag cloud with AV model for Topic 2, Topic 3, Topic 4, and Topic 6 (from left to right). For space reasons, the nomenclature can be found directly in the text.

features (hogv.entropy, hogv.mean) and pitch (mean.pitch, min.pitch), as shown in Figure 7. The drop in performance in terms of R^2 of Gloomy, Clumsy, and Charming cannot be explained on the basis of personality traits because Gloomy has only one likely personality related word. Instead, Gloomy is described by high arousal negative emotion words such as "disappointed", and "stressed", which could be associated to some acoustic or visual activity. Figure 7 mainly relates this topic to voicing rate (mean.val.apeak) and pitch (sd.nPcands) related features. Clumsy showed one main predictor associated to the distance to the camera (mean.face.size), and other cues such as not-looking-while-speaking and not-looking-while-not-speaking. With respect to the distance cue, we hypothesize that a certain proximity to the camera may not be very appealing to audiences, but this needs further exploration in future work. For the multimodal patterns, it seems plausible that some of these behaviors may create poor impressions of vloggers. Finally, Charming was mainly predicted from pitch features. This is explained with the topic being populated by females. Among the Top 10, 20, and 50 vloggers in Charming, we found 8, 18, and 43 females respectively.

2) Speech transcriptions: In general, topics not predicted with the AV model were found to be predictable using the TRA model. This concurs with previous literature showing that some personality traits are better predicted with verbal content, compared to nonverbal behavior [7]. In particular, the highest prediction with TRA was achieved for Geek (R^2 = .20) and Grouchy (R^2 = .19), Easygoing (R^2 = .11), and to much lesser extent for Gloomy (R^2 = .07). were predicted by TRA with lower performance. The performance of TRA for both Grouchy and Geek can be explained by the presence of Conscientiousness, Agreeableness and Emotional Stability among the topics' most likely words, which are the traits best predicted by speech transcriptions in previous work [7]. In particular, Grouchy in Figure 3 captured the low scores of these traits ("low-cons", and "low-agr", and "low-emot"), and Geek captured the high counterparts ("high-cons", and "high-agr", and "high-emot"). As with the AV model, Gloomy and Geek were the ones showing lower results.

3) Comment features: We present and discuss the experiments with YouTube comments in more detail and extension than previous subsections as, to our knowledge, this is the first time that social media comments are investigated for prediction and first impressions. Overall, the results using YouTube comments showed lower performance than using AV and TRA. Using COLIWC, best results were achieved for Grouchy (R^2 = .14), and Geek (R^2 = .13), and to lesser extent Easygoing (R^2 = .08) and Charming (R^2 = .07). Using CONGR, best

results were achieved with Gloomy ($R^2 = .10$), Easygoing ($R^2 = .08$) and Geek ($R^2 = .08$). With the exception of CONGR and Gloomy, the results using verbal content from comments show a similar trend to results using verbal content from transcriptions, specially if we compare the performance of Grouchy, Easygoing, and Geek with respect the rest of the topics in Figure 6. This is relevant, because at this point, it remains uncertain what information from comments can be potentially used by automatic models to predict impressions from vloggers. One possible explanation is that commenters directly provide impression information on their comments. Another explanation is that comments and transcripts contain similar information, which could be the case if people tend to comment on what vlogger said in the video.

We tested the second hypothesis using an information retrieval approach by measuring the similarity between transcripts and comment threads. To do so, we represented both types of documents with unigrams, and computed the cosine similarity matrix resulting from all pair-wise combinations of transcripts and comment threads. Then, we compared the distances between corresponding transcripts and comment thread pairs with respect to the rest of the threads. For each transcript, we ranked all comment threads by decreasing similarity and measured the retrieval precision as $p = 1/k$, where k is the position where the relevant document was found. Using this approach, very high precision indicates that a transcription and its corresponding comment thread are very similar compared to the rest of the threads. Figure 8 shows the average precision achieved with respect of the comment thread length (in number of comments). In addition to report the average precision $\hat{p} = \frac{\sum_i p_i}{N}$ (where N is the number of transcriptions), we also report the average pair-wise similarity between transcripts and corresponding comment threads ($\hat{s_p}$), and between transcripts and other comment threads ($\hat{s_o}$).

As shown in Figure 8 (left), the average precision when retrieving comment threads was low ($\hat{p} = .36$) and the similarity was moderate ($\hat{s_p} = .43$, $\hat{s_o} = .37$). We also observed that longer comment threads were more similar to transcripts, which could be explained by stopwords dominating the unigram representation. Interestingly, we found that removing stop words increased dramatically the precision to $\hat{p} = .69$ and that the difference between $\hat{s_p}$ and $\hat{s_o}$ increases substantially ($\hat{s_p} = .27$, $\hat{s_o} = .08$). This result supports the idea that documents and transcripts contain similar verbal content, and in part, explains why comments were found to be useful to classify the content of videos in a related work [18].

However, in spite of the similarity between comments and transcripts, the COLIWC and CONGR comment-based models did not provide results comparable to those of speech transcripts (TRA). For the latter, we argued that our experiments using n-grams suffer from not having enough data. For the former, we also recomputed the cosine similarity between comments and transcripts based on the LIWC representation to understand how much of the similarity between transcripts and comments hold. After processed by LIWC, Figure 8 (right) shows that the average precision drops with ($\hat{p} = .34$), and without stop words ($\hat{p} = .30$). However, we note that both $\hat{s_p}$ and $\hat{s_o}$ become very high in both cases ($\hat{s_p} = .94$, $\hat{s_o} = .93$, $\hat{s_p} = .63$,$\hat{s_o} = .68$), which could be due to the implicit dimensionality reduction of LIWC.

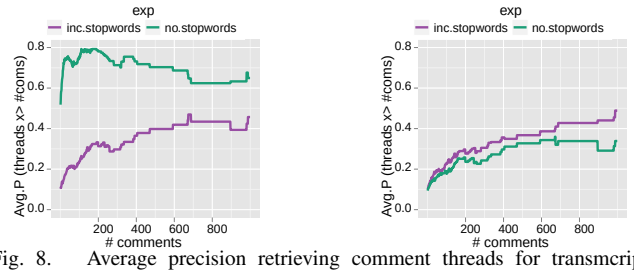

Fig. 8. Average precision retrieving comment threads for transmcripts: (left) Using unigrams; (right) using LIWC. Precision based on unigrams after removing stop words indicates similarity between transcripts and comment threads. For LIWC, precision falls due to the high similarity between all comment threads and transcripts once represented by LIWC categories (please view in color).

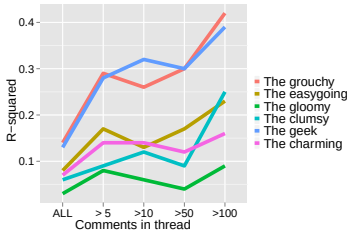

Fig. 9. Variation of the R-squared performance with the number of comments in thread.

One of the limitations of our experiments is the amount of data available. This has a clear effect for CONGR, as the number of features is larger than the number of samples used for training. However, there is also an effect of some comment threads being very short, which may result in poor representations. To test the effect of comment thread length in the prediction of topical impressions, we run experiments on different subsets of data by by training and testing models using comment thread lengths of 5, 10, 50, and 100 comments different subsets. As shown in Figure 9, the performance of the predictor increases substantially, specially for Grouchy (up to $R^2 = 42$) and Geek ($R^2 = 39$), which are the topics best predicted with the TRA model. Clearly, longer aggregated comments provide more information than shorter comments, and this may have implications when training the model.

4) Feature combinations: We finally explored all different combination of nonverbal, speech transcripts and comment features putting the three full features sets as one, and using the same training/test setting that with single modalities. As shown by the error bars in Fig. 6, model combination did not result in significative improvements with respect to best single models, but did not degrade either.

VII. Conclusions

We presented an original investigation on crowdsourced human impressions on a dataset of conversational vlogs, which includes three main contributions. First, our work adds to emerging research of interpersonal perception in social media by investigating impressions formation in vlogging beyond personality traits, with the inclusion of attractiveness, and mood.

Our analysis showed that MTurk annotators agree substantially on their impressions of Big-Five personality traits, attractiveness facets and moods, which supports previous literature showing the feasibility to crowdsource interpersonal human impressions. In this context, however, it is still unclear whether the low reliability measured for some traits, attractiveness facets, and moods is due to their poor manifestation in conversational vlogging, because they are difficult to annotate on the basis of thin-slices, or both. We also provided insights on the interplay among impressions, on the light of existing literature in psychology. Overall, we believe that the amount of behavioral data available in YouTube could help to back up other findings from social psychology at a scale not done before. Future work can study the validity of impressions compared to self-reported traits, specially for personality. Moreover, it would be interesting to study to what extent impressions from vloggers are stable throughout the longitudinal data available in vlog collections. Finally, future work may also explore the annotation of other personal dimensions different than the ones explored here, and that may be very relevant to conversational social video like persuasion.

Second, we investigated the use of a bag-of-impressions representation and LDA to mine multifaceted impressions in an unsupervised manner. Our experiments show that, indeed, the discovered topics result from combining personality, attractiveness, and mood facets, and that the topic model is useful to capture relationships that result from the interplay between these facets. In addition, we validated the quality of vlog-topic assignments found by the topic model through a crowdsourced topic intrusion task. Future work could explore the use of LDA or similar methods to mine an open (i.e. unrestricted) vocabulary of vlog first impressions. It may also investigate the generalization of the LDA output obtained here to a larger to larger sample of data. Overall, we argue that this type of topical characterization of vloggers, can be useful for tasks such as vlogger retrieval, which may be a relevant task in certain scenarios or future applications.

Finally, we addressed the prediction of topic-derived multifaceted impressions for the first time, going beyond the tasks of personality or mood prediction recently studied in the current social media literature. Our experiments showed that different sets of features: audiovisual analysis (AV), transcriptions (TRA), and YouTube comments (COLIWC), were useful to predict several topics with performances up to $R^2 = .30$. These results, while not directly comparable to previous literature due to the novelty of the task are encouraging because they are in the same range than performances achieved in pervious personality prediction tasks [8]. Our experiments also show that feature selection provide only substantial advantage to the use of the full feature sets. Future work may investigate this issue specially on what concerns the combination and selection of features from different modalities.

We also found that the performance of YouTube comments could be explained by the similarity between comment content and transcripts and that both TRA and COLIWC models were able to predict the same topics and shared some of the top predictors. We also observe that a drop in performance of COLIWC with respect to TRA could be explained by having very short comments, and that taking comment threads longer than 5 comments substantially increased performance, though, overall, these experiments need to be backed up with more data in future studies. The results with comments could have key implications because comments could potentially replace the automatic analysis of vlogger verbal content in some settings, given the current performance of automatic speech recognizers. Finally, we saw that combining feature sets achieved comparable performance to best single sets.

ACKNOWLEDGMENT

We thank the support of the Swiss National Science Foundation (SNSF) through the IM2 project.

REFERENCES

[1] N. Ambady, M. Hallahan, and R. Rosenthal, "On judging and being judged accurately in zero-acquaintance situations," *Journal of Personality and Social Psychology*, vol. 69, no. 3, pp. 518–528, 1995.

[2] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis," *Psychological Bulletin*, vol. 111, no. 2, pp. 256–274, 1992.

[3] K. Balog and M. de Rijke, "Decomposing bloggers' moods," in *Proc. of the Int. Conf. on World Wide Web (WWW)*, 2006.

[4] T. S. Behrend, D. J. Sharek, A. W. Meade, and E. N. Wiebe, "The viability of crowdsourcing for survey research," *Behavior research methods*, vol. 43, no. 3, pp. 800–813, 2011.

[5] A. J. Berinsky, G. A. Huber, and G. S. Lenz, "Using mechanical turk as a subject recruitment tool for experimental research," *Political Analysis 2*, 2012.

[6] J.-I. Biel and D. Gatica-Perez, "The good, the bad, and the angry: Analyzing crowdsourced impressions of vloggers," in *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM),*, 2012.

[7] J.-I. Biel, V. Vtsminaki, J. Dines, and D. Gatica-Perez, "Hi youtube! personality impressions and verbal content in social video." in *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2013.

[8] J.-I. Biel and D. Gatica-Perez, "The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 41–55, 2013.

[9] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez, "FaceTube: predicting personality from facial expressions of emotion in online conversational video," in *Proc. of Int. Conf. of Multimodal Interfaces (ICMI-MLMI)*, 2012.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[11] P. Borkenau and A. Liebler, "Trait inferences: Sources of validity at zero acquaintance," *Journal of Personality and Social Psychology*, no. 62, pp. 645–657, 1992.

[12] A. Brew, D. Greene, and P. Cunningham, "Using crowdsourcing and active learning to track sentiment in online media," in *Proc. of European Conf. on Artificial Intelligence (ECAI)*, 2010.

[13] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk," *Perspectives on Psychological Science*, vol. 6, no. 1, p. 3, 2011.

[14] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009, pp. 288–296.

[15] N. A. Diakopoulos and D. A. Shamma, "Characterizing debate performance via aggregated twitter sentiment," in *Proc. of ACM SIG Int. Conf. on Human factors in Computing Systems (CHI)*, 2010.

[16] K. K. Dion, A. W. Pak, and K. L. Dion, "Stereotyping physical attractiveness: A sociocultural perspective," vol. 21, no. 2, pp. 158–179, 1990.

[17] D. C. Evans, S. D. Gosling, and A. Carroll, "What elements of an online social networking profile predict target-rater agreement in personality impressions?" in *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2008.

[18] K. Filippova and K. B. Hall, "Improved video categorization from text metadata and user comments," in *Proc. of ACM SIGIR Int. Conf. on Res. and Develop. in Inf. Retrieval*, 2011.

[19] A. Fiore, L. Taylor, G. Mendelsohn, and M. Hearst, "Assessing attractiveness in online dating profiles," in *Proc. of ACM SIG Int. Conf. on Human factors in Computing Systems (CHI)*, 2008.

[20] S. D. Gosling, S. Gaddis, and S. Vazire, "Personality impressions based on Facebook profiles," in *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2007.

[21] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, "A very brief measure of the big five personality domains," *Journal of Research in Personality*, vol. 37, pp. 504–528, 2003.

[22] S. Gosling, S. Ko, and T. Mannarelli, "A room with a cue: Personality judgments based on offices and bedrooms." *Journal of Research in Personality*, vol. 82, pp. 379–98, 2002.

[23] G. Hitsch, A. Hortacsu, and D. Ariely, "What makes you click: An empirical analysis of online dating," in *Society for Economic Dynamics*, vol. 207, 2005.

[24] T. Iwata, T. Yamada, and N. Ueda, "Modeling social annotation data with content relevance using a topic model." in *NIPS*, vol. 9, 2009, pp. 835–843.

[25] K. M. Kniffin and D. S. Wilson, "The effect of nonphysical traits on the perception of physical attractiveness: Three naturalistic studies," *Evolution and Human Behavior*, vol. 25, no. 2, pp. 88 – 101, 2004.

[26] A. Kramer, "An unobtrusive behavioral model of gross national happiness," in *Proc. of ACM SIG Int. Conf. on Human factors in Computing Systems (CHI)*, 2010.

[27] R. Krestel, P. Fankhauser, and W. Nejdl, "Latent dirichlet allocation for tag recommendation," in *Proceedings of the third ACM conference on Recommender systems*. ACM, 2009, pp. 61–68.

[28] B. Lepri, N. Mana, A. Cappelletti, F. Pianesi, and M. Zancanaro, "Modeling the personality of participants during group interactions," in *Proc. of Int. Conf. on User Model. , Adapt. , and Personalization,*, 2009.

[29] J. Li and M. Chignell, "Birds of a feather: How personality influences blog writing and reading," *International Journal of Human-Computer Studies*, vol. 68, no. 9, pp. 586–602, 2010.

[30] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–501, 2007.

[31] R. R. McCrae and O. P. John., "An introduction to the five-factor model and its applications," *Journal of Psychology*, vol. 60, pp. 175–215, 1992.

[32] M. Mehl, S. Gosling, and J. Pennebaker, "Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life." *Jour. of Per. and Social Psych.*, vol. 90, no. 5, p. 862, 2006.

[33] R. Mihalcea and H. Liu, "A corpus-based approach to finding happiness," in *Proc. of Spring Symposia on Computational Approaches to Analyzing Weblogs (CAAW)*, 2006, p. 19.

[34] G. Mishne, "Experiments with mood classification in blog posts," in *Proc. of ACM SIGIR 2005 Stylistic Analysis Of Text For Information Access (SATIA)*, 2005.

[35] G. Mohammadi, A. Vinciarelli, and M. Mortillaro, "The voice of personality: Mapping nonverbal vocal behavior into trait attributions," in *Proc. of ACM Multimedia Workshop on Social Signal Processing (SSP)*, 2010.

[36] T. Nguyen, D. Phung, B. Adams, T. Tran, and S. Venkatesh, "Classification and pattern discovery of mood in weblogs," *Advances in Knowledge Discovery and Data Mining*, 2010.

[37] S. Nowson and J. Oberlander, "Identifying more bloggers: Towards large scale personality classification of personal weblogs," in *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2007.

[38] J. Pennebaker and L. King, "Linguistic styles: Language use as an individual difference." *Journal of Personality and Social Psychology*, vol. 77, no. 6, 1999.

[39] A. S. Pentland, *Honest Signals: How They Shape Our World*, ser. MIT Press Books. The MIT Press, 2008, vol. 1.

[40] P. Shrout and J. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, no. 2, p. 420–428, 1979.

[41] J. Snel, A. Tarasov, C. Cullen, and S. J. Delany, "A crowdsourcing approach to labelling a mood induced speech corpus," 2012.

[42] M. Soleymani and M. Larson, "Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus," in *Proc. of SIGIR Workshop on Crowdsourcing for Search Evaluation*, 2010.

[43] K. Stecher and S. Counts, "Spontaneous inference of personality traits and effects on memory for online profiles," in *Proc. AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2008.

[44] F. Steele Jr, D. C. Evans, and R. K. Green, "Is your profile picture worth 1000 words? Photo characteristics associated with personality impression agreement." in *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2009.

[45] D. J. Stillwell and M. Kosinski, "mypersonality project: Example of successful utilization of online social networks for large-scale social research," *American Psychologist*, vol. 59, no. 2, pp. 93–104, 2004.

[46] S. Vazire and S. D. Gosling, "e-Perceptions: Personality impressions based on personal websites," *Journal of Research in Personality*, vol. 87, pp. 123–132, 2004.

[47] J. B. Walther, B. Van Der Heide, S.-Y. Kim, D. Westerman, and S. T. Tong, "The role of friends' appearance and behavior on evaluations of individuals on facebook: Are we known by the company we keep?" *Human Communication Research*, vol. 34, no. 1, pp. 28–49, 2008.

[48] S. S. Wang, S. Moon, K. H. Kwon, C. A. Evans, and M. A. Stefanone, "Face off: Implications of visual cues on initiating friendship on facebook," *Computers in Human Behavior*, vol. 26, no. 2, pp. 226–234, 2010.

[49] L. Wu, L. Yang, N. Yu, and X.-S. Hua, "Learning to tag," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 361–370.

[50] T. Yarkoni, "Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers," *Journal of Research in Personality*, vol. 44, pp. 363–373, 2010.

Joan-Isaac Biel received his PhD from the Swiss Federal Institute of Technology in Lausanne (EPFL) in June 2013. He carried out his doctoral research at the Idiap Research Institute, and has visited Yahoo! Labs (Barcelona), HP Labs (Palo Alto), and the International Computer Science Institute (Berkeley). His research is focused on the analysis of human communication, interaction, and multimedia engagement with online social video.

Daniel Gatica-Perez is Head of the Social Computing Group at Idiap Research Institute and Matre d'Enseignement et de Recherche at the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland. His research interests include social computing, mobile and ubiquitous computing, and social media. He has served as Associate Editor of the IEEE Transactions on Multimedia. He is a member of the IEEE.