Hi YouTube! Personality Impressions and Verbal Content in Social Video

Joan-Isaac Biel[1,2],Vagia Tsiminaki [1], John Dines[1], and Daniel Gatica-Perez[1,2]
[1] Idiap Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
(jibiel,vtsiminaki,dines,gatica)@idiap.ch

ABSTRACT

Despite the evidence that social video conveys rich human personality information, research investigating the automatic prediction of personality impressions in vlogging has shown that, amongst the Big-Five traits, automatic nonverbal behavioral cues are useful to predict mainly the Extraversion trait. This finding, also reported in other conversational settings, indicates that personality information may be coded in other behavioral dimensions like the verbal channel, which has been less studied in multimodal interaction research. In this paper, we address the task of predicting personality impressions from vloggers based on what they say in their YouTube videos. First, we use manual transcripts of vlogs and verbal content analysis techniques to understand the ability of verbal content for the prediction of crowd-sourced Big-Five personality impressions. Second, we explore the feasibility of a fully-automatic framework in which transcripts are obtained using automatic speech recognition (ASR). Our results show that the analysis of error-free verbal content is useful to predict four of the Big-Five traits, three of them better than using nonverbal cues, and that the errors caused by the ASR system decrease the performance significantly.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human Factors

Keywords

YouTube, vlogs, verbal analysis, personality impressions

1. INTRODUCTION

The ubiquity of monologue videoblogs (i.e. vlogs) on the Internet has motivated social media researchers to investigate how interpersonal impressions are built based on the spontaneous behavior that vloggers display on their videos. Previous efforts analyzing YouTube vlogs have focused on identifying what nonverbal cues from audio and video correlate with Big-Five personality judgments of vloggers, and have also addressed the task of automatically predicting personality impressions [2]. However, despite the robustness of automatic nonverbal cue extraction and the reliability of personality judgments, these personality prediction experiments showed significant performance mainly for one of the Big-Five traits: Extraversion. Similar results were found investigating the potential of automatic facial emotion cues on the same task [3], which overall suggests that, given that people agree on their judgments of vloggers, personality information may be coded somewhere else, for example, in the verbal channel.

Social psychology has a long tradition of using text analysis in personality research to study the basic psychometric properties of word usage and to explore the relations between certain language dimensions and personality [19]. In social media, research has also exploited automatic text analysis to predict personality from large-scale text blog data [18, 17, 9], and from transcriptions of mobile SMS communication [8]. In this work, we investigate the feasibility of using verbal content for the prediction of personality impressions from vloggers using both manual speech transcriptions and automatic speech recognition (ASR) of YouTube conversational vlogs. Whereas the verbal channel is a clear alternative to the nonverbal channel, to our knowledge, this is the first time it is automatically extracted and analyzed in the context of conversational social video. In addition, ASR technologies are the only means to truly scale up verbal content analysis to the amount of online social video available today.

Our work has three contributions. First, we investigate the potential of using verbal content analysis in an error-free setting using manual transcriptions. We perform a standard cue utilization analysis to identify what aspects of nonverbal behavior are useful to build personality impressions from vloggers, and we put the focus in comparing our findings to previous research in text blogs. Second, we address the task of personality impression prediction from verbal content and investigate the feasibility of a fully-automatic framework by using a state-of-the-art ASR system. In addition, we explore two different approaches to model verbal content in vlogs proposed in the social media literature. Third, we benchmark the performance of verbal-based models with previous approaches to predict personality impressions using nonverbal cues measuring audiovisual activity and facial expressions of emotion. Amongst other results, our experiments show that verbal content can predict four of the Big-Five

ICMI'13, December 09 - 13 2013, Sydney, NSW, Australia
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ACM 978-1-4503-2129-7/13/12$15.00.
http://dx.doi.org/10.1145/2522848.2522894.

impressions, three of them much better than using nonverbal cues (which is only superior for Extraversion), that errors caused by the ASR system significantly degrade automatic predictions, and that combining nonverbal and verbal cues can increase prediction performance.

The rest of the paper is organized as follows. Section 2 reviews related work in social media and multimodal interaction. Section 3 overviews our approach to study verbal content and personality. Section 4 describes data and feature extraction methods. We discuss experiments and results in Section 5, and we conclude in Section 6.

2. RELATED WORK

Our study of personality impressions in social video relates to research investigating the utility of social media data to convey personality information. The problem has received special attention in the context of text blogging, were large-scale data has been used to back up earlier social psychology research linking individual personality differences and linguistic styles [19]. These works have studied verbal content usage from the perspective of both self-reported personality traits (i.e. how bloggers see themselves) [21, 4] and personality impressions (i.e. how readers see bloggers) [11], and have also been followed by several attempts to automatically predict personality from text blog data [18, 17, 9], and from manual transcriptions of mobile SMS communication [8]. Two main approaches to model verbal content were proposed in the above literature: a) using word category usage counts derived from a linguistic categorization system defined in social psychology research (the Linguistic Inquiry and Word Count) [19]; and b) using a standard data-driven representation based on n-gram frequencies. A systematic comparison of these two methods showed that, after an adequate feature selection, the n-gram based model outperformed LIWC on the prediction of blogger personality [9].

This work adds up to our previous research analyzing nonverbal behavior and personality impressions in the context of YouTube vlogs. In vlogging, video enables users to enrich their narratives with spontaneous nonverbal behavior through voice, gestures, pose, and face, and this becomes a potential source of personality information, often more difficult to control than verbal content. Previously, we investigated the feasibility of crowdsourcing personality impressions and addressed the task of automatically predicting personality impressions by focusing on the nonverbal aspect of vloggers' behavior [2, 3]. In a first approach that computed automatic nonverbal cues from audio and video [2], Big-Five trait impression prediction experiments showed substantial performance for the Extraversion trait (up to $R^2 = 36\%$), low performance for Openness to Experience and Conscientiousness (up to $R^2 = 10\%$ in both cases), and no significance performance for Agreeableness and Emotional Stability. In a second study investigating the potential of automatically extracted facial emotion cues [3], results were consistent on that Extraversion was the best trait predicted (up to $R^2 = 22\%$), and showed low performance for Openness to Experience (up to $R^2 = 12\%$), Agreeableness (up to $R^2 = 8\%$), and Conscientiousness (up to $R^2 = 7\%$). These prediction results contrasted with the reliabilities shown by crowdsourced personality impressions, which indicate that vlogs convey rich human personality information, and suggest that with the exception of Extraversion, useful cues for personality inference may be found in other aspects of behavior, for instance, in the verbal channel, which we investigate in this paper.

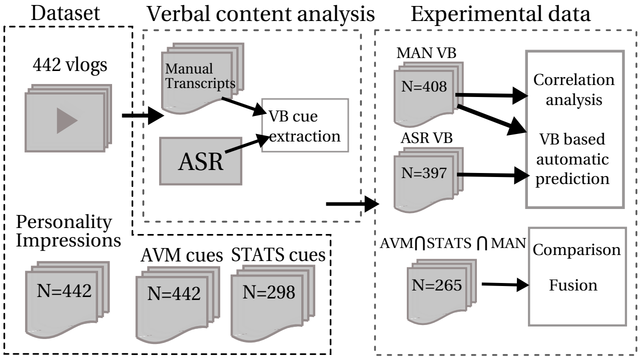

Figure 1: Our approach to investigate verbal content for vlogger personality impression prediction (AVM = audiovisual, STATS = facial cues, MAN = manual, VB = verbal, ASR = automatic speech recognition).

The discussion on what aspects of behavior convey personality information is also relevant to previous multimodal interaction research studying personality from a nonverbal perspective. Most of this literature has shown the potential of automatic nonverbal cues to predict the Extraversion trait. The automatic classification of personality impressions in professional radio broadcasts [16] found that Extraversion and Conscientiousness were the personality impressions with highest classification performance (with up to 73% and 71% accuracy) and that the rest of the traits were more difficult to predict. In face-to-face meetings [10], automatic nonverbal analysis predicted Extraversion with up to $R^2 = 22\%$ (no other Big-Five personality traits were investigated), whereas in self-presentations videos [1], research found that nonverbal behavior was useful to predict Conscientiousness, Emotional Stability and Extraversion (in this order). In this context, the only work we know of that has investigated verbal content and personality in face-to-face interactions is [15], were models using text features achieved better performance than models based on audio features for Extraversion ($R^2 = 24\%$), and Conscientiousness ($R^2 = 18\%$). Overall, the above literature shows that the effectiveness of nonverbal cues to predict traits other than Extraversion depends on the conversational setting and the prediction task (i.e. self-personality [10, 1, 15] vs. personality impressions [16, 15]). Based on social psychology literature [5], we hypothesize that part of the limited performance of these predictions comes from the limited ability of nonverbal cues to convey information for some traits.

Finally, regarding the use of ASR technologies to model verbal content, our work relates to a recent study addressing the task of automatically predicting emergent leadership in face-to-face conversations using verbal cues from automatic transcriptions [20].

3. OUR APPROACH

Our approach to investigate the value of verbal content analysis for automatic personality impression prediction in vlogs is summarized in Figure 1. We use a dataset of vlogs that includes videos, personality impressions, and two sets of nonverbal cues that were used in our previous work [2, 3]. These feature sets (described in Section 4.4) are used to benchmark verbal content predictions, but as shown in the figure, the automatic processing of vlogs in this paper is focused on the extraction of verbal cues from both manual and automatic transcripts.

The paper is composed presents three experiments. First, we frame the problem of cue utilization based on manual transcripts. In particular, we put the discussion in the context of previous research in verbal content and personality in text blogs. Second, we address the task of personality prediction based only on verbal content cues extracted form manual and automatic transcriptions. The manual transcripts are used to explore verbal content in an error-free setting, whereas the ASR output is used to investigate the feasibility of scaling up verbal content analysis to large amounts of vlog data using a fully-automatic framework. Third, we compare the prediction performance of verbal cues to the one obtained using audiovisual activity and facial expression cues, and we explore improvements when combining different modalities. Note that these last experiments are performed in a smaller dataset, which is the intersection of the features available for the dataset (audiovisual and facial cues) and the verbal cues extracted in this paper.

4. DATASET AND PROCESSING

The dataset used in this work was previously used in [2, 3], and consists of 442 YouTube vlogs and a collection of personality impressions for each vlogger.

The videos in this dataset show one single vlogger in front of a webcam talking about a variety of topics including personal issues, politics, movies, books, etc. We did not use any content-related restriction during data collection, so the language used in the videos is natural and diverse. The collection is mostly mostly balanced in gender, with 208 males (47%) and 234 females (53%).

The personality impressions consist of Big-Five personality scores [14] that were collected throughout a crowdsourcing annotation task [2]. In this task, annotators watched one-minute slices of each vlog, and rated impressions using a personality questionnaire. As reported in [2], the aggregated Big-Five scores are reliable with the following intra-class correlations (ICC(1,k), k=5): Extraversion ($ICC = .76$), Agreeableness ($ICC = .64$), Conscientiousness ($ICC = .45$), Emotional Stability ($ICC = .42$), Openness to Experience ($ICC = .47$), all significant with $p < 10^{-3}$. Note that whereas Extraversion has been typically found easy to judge in many scenarios, the high reliability of Agreeableness seems particular to the vlogging setting. However, despite the agreement achieved on judging this trait, our previous research on predicting personality using nonverbal cues [2, 3] mainly showed significant results for Extraversion. Check [2] for further details on the crowdsourcing task.

4.1 Manual Transcriptions

We asked a professional company to manually transcribe the audio from vlogs. The data sent for annotation corresponded to the full vlog duration from 426 randomly selected vloggers, up to a total of 30h of audio. 5% of the vloggers were later discarded during the transcription process because their audio was unintelligible or they were not speaking English.

Word Recognition Performance	
Correct (C)	45.6%
Substitution (S)	28.5%
Deletions (D)	25.9%
Insertions (I)	8.0%
Total Error (WER)	62.4%

Table 1: Word automatic recognition performance in 397 vlogs compared to automatically aligned manual transcriptions. $WER = \frac{S+D+I}{S+D+C}$.

The resulting manual transcripts comprise 408 vloggers out of the 442 with personality impressions, and they are also balanced in gender: 197 males (48%) and 211 females (52%). The whole transcriptions contained a total of 10K unique words and 243K word tokens, and do not include timestamps.

4.2 Automatic Transcriptions

We used a state-of-the-art automatic speech recognizer to generate transcriptions for the audio files and to align the manual transcriptions (used later for evaluation). The system combines two two-pass English systems that use acoustic models based on individual head-mounted microphones (IHM) and single-distant microphone (SDM), respectively. Both the IHM and SDM systems use identical decoding configurations, e.g., the first pass uses unadapted acoustic models, followed by unsupervised adaptation in the second pass. The four hypotheses from both the first and second passes of IHM and MDM are aligned and combined to produce word-level confidence scores, and decoded with a weighted finite-state transducer, using a lexicon of 50,000 words and a 4-gram language model trained on various corpora for a total amount of about one billion words. More details can be found in [6].

We run the system on the 408 vlogs with manual transcriptions, but failed in up to 11 vlogs during alignment or decoding mostly because of the low audio quality and background noise. As summarized in Table 1, the ASR system achieved a word error rate (WER) of 62.4%, with respect to the aligned manual transcriptions for the 397 successfully decoded files. Note that part of this WER may be due to misalignments of the manual transcriptions, though the actual percentage is unknown.

Though these results may seem low compared to the WER achieved in other domains, they concur well with another recent work that automatically transcribed YouTube videos [7]. In those experiments, an ASR system trained on acoustic models from 1,400h of aligned YouTube audio incremented by more than double the WER of more controlled experiments, achieving up to a WER of 52.3%. These results clearly illustrate the current difficulty of automatically transcribing online social video.

4.3 Automatic Analysis of Verbal Content

We explore two methods to compute verbal content cues that have been previously used in social media to analyze text from blogs, and that we describe as follows.

4.3.1 LIWC

The first approach models verbal content on the basis of lexical features computed with the Linguistic Inquiry and

Word Count (LIWC) software [19]. This tool is the result of years of social psychology research focused on validating the psychometric properties of a word categorization system that links **linguistic and paralinguistic categories** to psychological constructs and personal concerns. In its English version, LIWC is built based on an dictionary composed of 4,500 words and word stems. Each word in the vlog transcript is looked up in the dictionary, and in case of a match the appropriate word category is incremented (note that in LIWC, words can be assigned to more than one category at a time). After a document is processed, the counts are divided by the total number of words in the document.

In this paper, we consider a total of 65 LIWC cues: we discarded the 12 punctuation categories, as there are not relevant in the spoken setting, and we included only one general descriptor that counts the words longer than six letters. Since LIWC is designed to process raw text, there is no need for any type of preprocessing.

4.3.2 N-grams

The second approach is a data-driven representation of verbal content based on n-grams, which is a standard model of text used in many tasks related to information retrieval and document classification. In particular, we replicated the approach in [9], where Weka's Correlation-based Feature Subset Selection (CFS) is used to select significant features for each prediction task using all available data. This feature selection step was shown to be key to outperform LIWC with n-grams in [9]. However, we believe that this selection procedure is prone to overfitting, and therefore, we consider applying CFS in two different settings to contrast the results. In the first setting, we use CFS inside the evaluation set up (during model training) to select features using only the training data, whereas in the second, we use CFS outside the evaluation setup with all the available data, as in [9]. In our experiments, we refer to these settings as inCFS and outCFS respectively.

Prior to generating n-grams, we preprocessed text by stemming words using Porter's stemming algorithm, removing punctuation, and omitting words that appeared in less than ten documents or less than ten times in the whole collection (we do not remove stop words). Then, we processed text to generate **unigrams and bigrams** and to compute $tf \cdot idf$ values for each n-gram.

Table 2 summarizes the amount of manual and automatic transcription data, including raw data (words), and the LIWC and n-gram outputs. In average, 91.7 % of the words from manual transcripts were found in the LIWC dictionary, a percentage that decreases to 66% for automatic transcripts. The actual feature sets for unigram and bigrams are much smaller after using inCFS and outCFS (in most cases they included no more than 100 features).

4.4 Nonverbal cues

Two different sets of nonverbal cues were gathered from our previous works in [2, 3] (i.e. they were not extracted in the framework of this paper). The first set is composed of **audio, visual and multimodal nonverbal cues** thereby referred to as AVM) that we introduced in [2]. From audio, it includes 3 speaking activity features measured on the basis of speech-non-speech segmentations (e.g. 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 video, it includes 3 looking activity cues from looking-non-looking (e.g. 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 (wMEI) that measure the accumulated amount of motion through the video. Finally, it also includes 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 130 cues that were available for 442 videos.

	Manual				Automatic	
	Words	LIWC	Uni	Bi	Words	LIWC
# Terms	10K	65	1K	287	7,6K	65
# Tokens	246K	221K	241K	110K	152K	142K

Table 2: Number of unique terms and tokens in manual and automatic data: raw vocabulary (words) and data processed using LIWC and n-grams (Uni, Bi).

The second set captures nonverbal cues from **facial expressions of emotion** and was used generated for our work in [3]. It is composed of 5 statistical aggregates extracted from frame-by-frame estimates of 7 facial basic expressions, one neutral signal, and smile (hereafter referred to as STATS) generated with the Computer Expression Recognition Toolbox (CERT) [12]. Summing up, these are 45 features available for a subset of 298 vlogs (the details on selecting this data sample are explained in [3]). The intersect between vlogs with audiovisual features, vlogs with facial cues, and vlogs with verbal content cues is of 265 videos (see Fig 1).

5. EXPERIMENTS AND RESULTS

We divided our experiments in two sections. Section 5.1 addressed the cue utilization analysis, whereas Section 5.2 presents the automatic prediction experiments.

5.1 Correlation Analysis

We computed pair-wise Pearson's correlations between the 66 word categories and personality traits to analyze the level of cue utilization with personality impressions. We did not perform this analysis for n-grams features because of the sparse representation, but also because related literature studying cue utilization has focused on the LIWC representation.

Table 3 summarizes the significant correlations ($p < .05$) between LIWC categories and Big-Five scores (from most negative to most positive). We found a total of 12 significant effects for judgments of Extraversion, 8 out of which are backed up by the literature. For example, we found that Extraversion judgments were associated with an increased use of categories related to interpersonal interaction (*you*, $r = .13$, *social*: $r = .10$) [19, 4]. As in [4], we found that Extraversion is the only trait associated with the use of 2nd person of singular (i.e., vloggers refer frequently to the YouTube audience). The increased use of sexual words (*sexual*, $r = .17$,) associated to the Extraversion trait is also documented in previous work [21]. We also found that vloggers judged as introverted use more cognitive related words (*cogmech*: $r = -.14$) , including discrepancy (*dis-*

Trait	LIWC categories
Extr	tentat (-0.19^{***}), nonfl (-0.18^{**}), cogmech (-0.14^{*}), discrep (-0.13^{*}), excl ($-0.12^{\dagger}$), ipron ($-0.11^{\dagger}$), health ($-0.10^{\dagger}$), social ($0.10^{\dagger}$), affect ($0.11^{\dagger}$), assent (0.13^{*}), you (0.13^{*}), space (0.16^{**}), sexual (0.17^{**}) Cue utilization = 12
Cons	assent (-0.23^{***}), i (-0.23^{***}), filler (-0.22^{***}), negemo (-0.19^{***}), ppron (-0.19^{***}), negate (-0.18^{**}), verb (-0.17^{**}), anger (-0.17^{**}), pronoun (-0.15^{*}), present (-0.15^{*}), swear (-0.15^{*}), adverb (-0.14^{*}), sexual (-0.14^{*}), auxverb (-0.14^{*}), time (-0.13^{*}), body ($-0.10^{\dagger}$), they ($0.11^{\dagger}$), discrep ($0.12^{\dagger}$), article (0.12^{*}), incl (0.17^{**}), achieve (0.17^{**}), work (0.21^{***}), preps (0.24^{***}), Sixltr (0.25^{***}) Cue utilization = 24
Open	health (-0.14^{*}), anger (-0.13^{*}), negemo (-0.12^{*}), nonfl ($-0.12^{\dagger}$), sad ($-0.11^{\dagger}$), swear ($-0.10^{\dagger}$), death ($-0.10^{\dagger}$), hear ($0.10^{\dagger}$), motion ($0.10^{\dagger}$), leisure (0.13^{*}) Cue utilization = 10
Agr	anger (-0.43^{***}), negemo (-0.42^{***}), swear (-0.37^{***}), sexual (-0.28^{***}), bio (-0.17^{**}), negate (-0.14^{*}), relig (-0.14^{*}), they (-0.13^{*}), quant ($-0.11^{\dagger}$), work ($0.09^{\dagger}$), friend ($0.11^{\dagger}$), incl ($0.12^{\dagger}$), i ($0.12^{\dagger}$), conj (0.14^{*}), posemo (0.24^{***}) Cue utilization = 15
Emot	negemo (-0.38^{***}), anger (-0.34^{***}), swear (-0.31^{***}), sexual (-0.31^{***}), bio (-0.17^{**}), negate (-0.16^{**}), affect ($-0.10^{\dagger}$), nonfl ($0.09^{\dagger}$), discrep ($0.10^{\dagger}$), work ($0.12^{\dagger}$), leisure ($0.12^{\dagger}$), achieve (0.12^{*}) Cue utilization = 12

Table 3: Selection of significant Pearson's correlation effects (p <.05) between LIWC cues personality impressions, ($^{\dagger}p < .05,^{*}p < .01,^{**}p < .001,^{***}p < .0001$).

crep: $r = -.13$), tentative (*tentat*: $r = -.19$), and exclusive words (*excl*: $r = -.12$), concurring with previous literature [21]. In addition, as in face-to-face interactions [15], we found Extraversion judgments associated with the expression of emotions (*affect*: $r = .11$).

Conscientiousness judgments showed 24 significant effects. Not surprisingly, we found Conscientiousness impressions to show some of the largest associations with words related to occupation and achievement (*work*: $r = .21$, *achieve*: $r = .17$), which is consistent with findings that associate Conscientiousness to an increase usage of these word categories [4]. These vloggers are also associated to a decreased use of negative emotion words (*negate*: $r = -.18$, *negemo*: $r = -.19$) [15], swearing words (*swear*: $r = -.15$), and sexual words (*sexual*: $r = -.14$) [21]. Though we did not find any positive association between Conscientiousness and the 3rd person pronoun as in [4], we found a negative association on the use of the 1st person pronoun (*i*: $r = -.23$). Other effects, not documented in the literature, are the correlation with auxiliary verbs (*auxverb*: $r = -.14$) and present tense (*present*: $r = -.15$), and the positive association with prepositions (*preps*: $r = .24$), articles (*article*: $r = .12$), and inclusive words (*incl*: $r = .17$). We also found that this trait was the only one positively correlated with the length of the words (*Sixltr*, $r = .25$), an effect that is associated to a careful choice of words [15]

We found 10 effects for Openness to Experience impressions. Vloggers judged as open to experience tend to use more words related to leisure activities (*leisure*: $r = .13$) [4], and words concerning senses (hear : $r = .13$) [4]. In addition, they tend to express negative emotions less frequently (*anger*: $r = -.13$, *negemo*: $r = -.15$, *anger*: $r = -.13$) [19, 4].

Agreeableness impressions displayed significant effects with 15 LIWC categories and sub-categories. First, as shown in previous literature [19], we found the largest effects for Agreeableness judgments with the use of both positive (*posemo*: $r = .24$), and negative emotions (*anger*: $r = -.43$, *negemo*: $r = -.42$). As discussed in [15], these word categories are prominently used by socially oriented and unconfrontational people. This relates to the idea that agreeable people are socially oriented and tend to avoid conflict [15]. Also concurring with previous work, we found positive associations with the use of self references (*i*: $r = .12$) [21], friendship (*friend*: $r = .11$) [15], and a negative association with *they* ($r = -.13$). In contrast, the large correlations with *anger* ($r = -.43$) and *negemo* ($r = -.42$) show that annotators associated the use of these type of words with less agreeable people. In addition, less agreeable people also use more words related to sexuality (*sexual*: $r = -.29$), swear words (*swear*: $r = -.37$), body states (*bio*: $r = -.17$), and religion (*relig*: $r = -.14$).

Finally, Emotional Stability showed 12 significant effects. High emotional scorers of this trait are negatively associated to the expression of negative words (*negate*: $-.16$), negative emotional words (*anger*: $r = -.34$, *negemo*: $r = -.30$, *affect*: $r = -.10$) [19], swear words (*swear*: $-.31$) and sexual words (*sexual*: $-.31$).

To conclude, it is important to note that despite that these correlation values may seem relatively low, they are within the order of magnitude reported in previous research analyzing verbal content automatically [21].

5.2 Automatic Personality Impression Prediction

We treated personality inference as five independent regression problems intended to predict the personality scores for each of the Big-Five impressions. Compared to other prediction tasks proposed in the literature such as personality classification or ranking, the regression task is the one that provides the most fine-grained personality recognition assessment [13]. For each task, we evaluated the use of two supervised machine learning predictors: Support Vector Machines (SVMs) with linear, polynomial, and RBF kernels, and Random Forests (RFs). We conducted experiments using a 10-fold cross-validation (CV): at each resample iteration, we trained a model using 9 folds of data, and test it on the left-out fold. To optimize model parameters, we used 5-fold cross validation on the 9 folds used for training. Note that we use CV with RF for practical reasons, but results were the same using the out-of-bag estimates of RF (i.e., the performance estimates computed on bootstrap left-out data). Whereas the linear kernel consistently underperformed the RBF and the polynomial kernel, the performance of the RBF and the polynomial kernel was almost the same for all the tasks (only in few cases the RBF provided slightly better performance than the polynomial). Hence, to keep the presentation of the results concise, we only report performance for the SVMs using RBF kernel, as well as for RFs.

In all the experiments, we measured the performance of automatic predictions using the coefficient of determination

Features	Extr	Cons	Open	Agr	Emot
Baseline	.00	.00	.00	.00	.00
$SVMRadial$					
LIWC	.14* (.12)	.19** (.11)	.04† (.04)	.26** (.13)	.08*(.07)
Unigrams	.05† (.05)	.14** (.08)	.03† (.06)	.14* (.09)	.07*(.05)
Bigrams	.04† (.05)	.08* (.06)	.03* (.02)	.13* (.12)	.04†(.04)
RF					
LIWC	.13* (.10)	.18** (.10)	.04**(.02)	.31***(.12)	.17*(.13)
Unigrams	.11***(.04)	.14***(.05)	.03† (.03)	.21** (.13)	.12*(.11)
Bigrams	.04† (.04)	.12** (.06)	.02† (.04)	.14* (.11)	.11†(.10)
Highest achieved performance					
	.14	.19	.04	.31	.17

Table 4: R-squared results on predicting personality impressions using SVM and RF for LIWC, unigram, and bigram cues computed in manual speech transcriptions, ($^{\dagger}p < .05,^{*}p < .01,^{**}p < .001,^{***}p < .0001$).

Features	Extr	Cons	Open	Agr	Emot
Baseline	.00	.00	.00	.00	.00
$SVMRadial$					
Uni-inCFS	.04* (.03)	.06* (.06)	.00† (.00)	.10** (.06)	.02† (.02)
Bi-inCFS	.02† (.03)	.01† (.01)	.00† (.00)	.05* (.04)	.04† (.04)
Uni-outCFS	.15***(.06)	.12* (.08)	.06** (.03)	.18** (.10)	.08* (.06)
Bi-outCFS	.24** (.12)	.14** (.08)	.16***(.07)	.17** (.09)	.13* (.10)
RF					
Uni-inCFS	.07* (.05)	.10* (.09)	.01* (.01)	.21***(.09)	.07† (.07)
Bi-inCFS	.03† (.03)	.04* (.03)	.00† (.01)	.08* (.05)	.11† (.12)
Uni-outCFS	.28***(.06)	.31***(.11)	.16** (.09)	.37***(.09)	.26**(.17)
Bi-outCFS	.33***(.13)	.32** (.16)	.28***(.08)	.41***(.14)	.32**(.16)
Highest achieved performance					
	.33	.32	.28	.41	.32

Table 5: R-squared results on predicting personality impressions using SVM and RF using unigrams (uni) and bigram (bi) with CFS inside cross-validation (inCFS) and outside (outCFS). Improvements for outCFS suggest overfitting, ($^{\dagger}p < .05,^{*}p < .01,^{**}p < .001,^{***}p < .0001$).

(R^2), as it the most frequently reported performance measured in personality regression [13, 10]. With this measure, the baseline regressor is a model that predicts the mean personality score (MPS) of the training data. To measure significant differences between the models and the baseline, we conducted single T-tests.

5.2.1 Results using manual transcripts

Table 4 summarizes the performance on the prediction of personality impressions using verbal content from manual transcriptions of vlogs (N=408). At a glance, we found that four of the Big-Five personality impressions can be predicted substantially better than the baseline. Though statistically significant, Openness to Experience predictions are close to the baseline. We also found that RF provided comparable or substantially higher performance than SVMs.

Results indicate that Agreeableness is the trait that shows higher performance ($R^2 = 31\%$), followed by Conscientiousness ($R^2 = .19$), despite the first trait having shown lower cue utilization than the latter. This result can be explained, in part, by the higher ICC reliability of Agreeableness impressions used to train the supervised models. It can also be explained by some LIWC categories including counts of other sub-categories. Finally, Emotional Stability and Extraversion achieved $R^2 = .17$ and $R^2 = .14$, respectively.

Table 4 also shows that LIWC cues provided superior performance than unigrams and bigrams using the inCFS setting. The comparison between using CFS inside (inCFS) and outside (outCFS) the cross-validation procedure can be found in Table 5. The difference between the performance obtained with the two methods indicates that using feature selection outside the training loop (as in [9]) may result in overfitting issues. Note that the improvement obtained using outCFS is larger for the case of the bigrams, an effect also seen in [9]). This occurs because bigrams tend to be sparser than unigrams, and therefore more prone to overfit small amounts of data. In addition, the performance of the bigram models is surprisingly high independently of the different reliability of impressions, which adds to the argument of overfitting. While it is clear that we cannot trust the outCFS procedure in our experiments, this result needs to be backed up with more data, which is limited here.

5.2.2 Results using automatic transcripts

Table 6 shows the results when using automatic transcriptions (N=397). Results are shown only for LIWC cues and the RF predictor, which was shown to be the best setting using manual transcripts. Results show that prediction performance decreases significantly when using automatic transcripts as a consequence of the errors introduced by the ASR

Features	Extr	Cons	Open	Agr	Emot
Base	.00	.00	.00	.00	.00
			RF		
liwc	$.02^{**}(.02)$	$.08^{**}(.06)$	$.02^{*}(.02)$	$.10^{**}(.08)$	$.05^{**}(.04)$
liwc-lowWER	$.04^{*}$ (.07)	$.10^{*}$ (.11)	$.05^{*}(.05)$	$.18^{**}(.12)$	$.10^{*}$ (.12)
liwc-highWER	$.07^{*}$ (.07)	$.10^{**}(.08)$	$.01^{*}(.02)$	$.12^{*}$ (.14)	$.05^{**}(.04)$
			Highest achieved performance		
	.07	.10	.05	.18	.10

Table 6: R-squared results on predicting personality impressions using and RF and LIWC for automatic transcriptions. liwc-lowWER and liwc-highWER are models retrained using automatic transcripts with low WER and high WER, respectively, ($^{\dagger}p < .05, ^{*}p < .01, ^{**}p < .001, ^{***}p < .0001$).

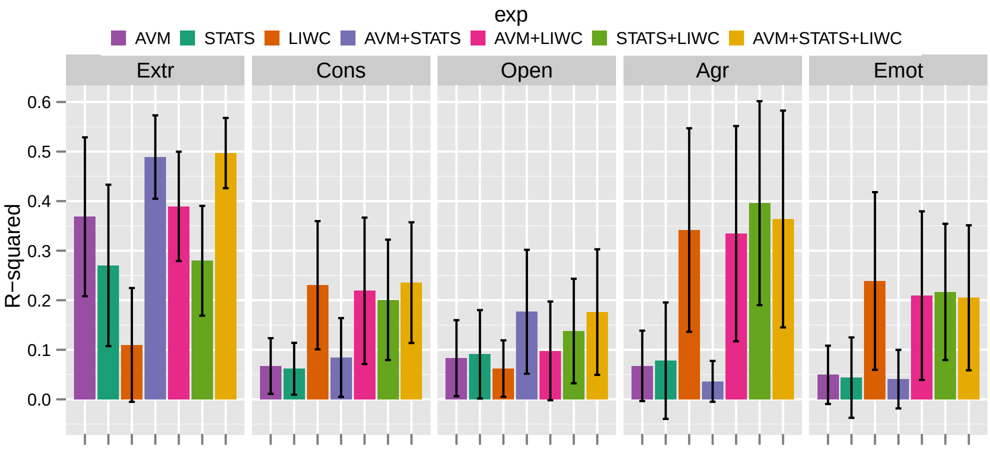

Figure 2: R-squared results on predicting personality impressions using RFs, best models for each modality (AVM for audiovisual, STATS for facial cues, and LIWC for verbal content), and combinations of them.

system. In particular, we see a drop in performance from $R^2 = .31$ to .10, for Agreeableness, and from $R^2 = .18$ to .08 for Conscientiousness, whereas for the rest of the traits the performance is close to the baseline. The effect of low automatic recognition performance on the prediction of personality impressions is evident when retraining two models using automatic transcripts with high and low WER. Compared to using all the data, the performance of the Agreeableness trait doubled for the subset of samples with low WER, despite the fact that the WER was still high ($WER = 77\%$). The results for Emotional Stability also improved, but for the rest of traits remain the same.

5.2.3 NVB and VB content: comparison and fusion

Figure 2 shows the prediction performance for the Big-Five traits using audiovisual nonverbal cues (AVM), facial expressions (STATs), and verbal content cues from manual transcriptions using LIWC (LIWC). The first three columns correspond to the single sets, and the four others result from combinations of them.

Note that though the dataset is smaller ($N = 265$), the results shown in this figure for AVM and STATs are consistent with results reported in [2] and [3]. For LIWC, results are marginally higher than the ones presented in the previous section. Best performance for audiovisual (AVM) and facial expressions cues (STATS) was achieved for Extraversion with $R^2 = .36$ ($p < 10^{-4}$) and $R^2 = .27$ ($p < 10^{-3}$) respectively. These results are consistent with previous research showing that Extraversion is the most predictable trait using nonverbal behavior, and also concur with the fact that it is the trait with highest impression reliability in the data used in our experiments. The figure also shows that using nonverbal cues for the other traits does not bring better performance compared to the baseline. Compared to nonverbal cues, verbal content is useful to predict Agreeableness ($R^2 = .34$, $p < 10^{-3}$), Emotional Stability ($R^2 = .24$, $p < 10^{-3}$), and Conscientiousness ($R^2 = .23$, $p < 10^{-3}$), and with lower performance, Extraversion ($R^2 = .11$, $p < 10^{-3}$). The result is specially relevant for Agreeableness, that despite having the second highest reliability is very difficult to predict with nonverbal cues, and overall, it indicates that personality information to make impressions of this trait comes mainly from the verbal modality. Similar results on the superior performance of verbal content to predict these traits was found on analyzing manual transcripts of face-to-face interactions in [13].

Experiments combining features also unveil some interesting findings. In particular, we found that two combinations of features substantially help to improve the performance for Extraversion, Agreeableness, and Openness to Experience. Combining audiovisual features and facial expressions (AVM+STATS) boosts the performance of Extraversion predictions up to $R^2 = .48$ ($p < 10^{-4}$), while for Openness to Experience, it doubles the performance of any single best predictor up to $R^2 = 17$ ($p < 10^{-2}$). In both cases, adding verbal content contributes to marginal improvements. In contrast, for Agreeableness, the performance improved when combining facial expression cues and verbal features (STATS+LIWC), from $R = .34$ ($p < 10^{-3}$) to $R = .39$ ($p < 10^{-3}$), whereas using the AVM cues did not contribute much.

6. CONCLUSIONS

While it is clear that automatic nonverbal cues convey useful information for the prediction of human personality, verbal content cues have been unexplored in many conversational scenarios, in part due to the cost of transcribing data. Nevertheless, the study of verbal content analysis in these settings is important to understand to what extent personality impressions are built on nonverbal behavior (i.e. how individuals say things) as opposed to verbal content (i.e. what they say), and can help assess the magnitude of the personality prediction performance that can be expected using current approaches.

In this paper, we investigated the value of verbal content for the prediction of vloggers' personality impressions. Our study using manual transcriptions concurs with previous research in text blog data on the type of significant correlation effects shown between verbal cues and personality traits. In addition, our experiments show that verbal content-based models can achieve better prediction performance than using nonverbal cues for three of the Big-Five traits. In particular, we found that Agreeableness was the trait with highest performance ($R^2 = .31$), followed by Conscientiousness ($R^2 = .19$) and Openness to Experience ($R^2 = .17$). Specially for the case of Agreeableness, the result is important because despite being the second trait with higher reliability, the performance obtained by audiovisual activity and facial cues models was poor.

Our experiments also showed that the performance of verbal content models decreases significantly when using automatic transcriptions due to errors introduced by the ASR system. Future work may exploit the ASR output confidences to filter out unreliable verbal content or use automatic keyword spotting instead of full ASR. However, it may just be that that we need ASR technologies to improve before we can start using them for this task. Finally, our work showed that the combination of different modalities, namely audiovisual activity and facial emotion cues for Extraversion, and facial and verbal cues for Agreeableness can improve prediction performance.

To conclude, we acknowledge that our experiments would benefit of having more data. In particular, this could help in experiments with n-grams, where the dimensionality of the feature vectors was considerably larger than the number of documents; and in fusion, to attempt other fusion methodologies.

Acknowledgments

This research was funded by the Swiss NSF through the NCCR IM2 (VlogSense and YouBlog projects).

7. REFERENCES

[1] L. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe. Please, tell me about yourself: Automatic assessment using short self-presentations. In *Proc. of Int. Conf. of Multimodal Interfaces (ICMI-MLMI)*, 2011.

[2] J.-I. Biel and D. Gatica-Perez. The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Trans. on Multimedia*, 15(1):41–55, 2013.

[3] 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 ACM ICMI*, 2012.

[4] A. J. Gill, S. Nowson, and J. Oberlander. What are they blogging about? Personality, topic and motivation in blogs. In *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2009.

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

[6] T. Hain et al. Transcribing meetings with the amida systems. *IEEE Trans. Audio, Speech and Lang. Proc.*, 20(2):486–498, Feb. 2012.

[7] Hinton et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 2012.

[8] T. Holtgraves. Text messaging, personality, and the social context. *Jour. of Res. in Pers.*, 45(1):92–99, 2011.

[9] F. Iacobelli, A. Gill, S. Nowson, and J. Oberlander. Large scale personality classification of bloggers. In *Proc of ACII*, 2011.

[10] 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 Modeling, Adaptation, and Personalization*, 2009.

[11] J. Li and M. Chignell. Birds of a feather: How personality influences blog writing and reading. *Int. Jour. of Human-Comp. Studies*, 68(9):589–602, 2010.

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

[13] 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*, 30:457–501, 2007.

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

[15] 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.*, 90(5):862, 2006.

[16] 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.

[17] 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.

[18] J. Oberlander and S. Nowson. Whose thumb is it anyway? Classifying author personality from weblog text. In *Proc. the Annual Meeting of the Assoc. for Computational Linguistics*, 2006.

[19] J. Pennebaker and L. King. Linguistic styles: Language use as an individual difference. *Jour. of Per. and Social Psych.*, 77(6):1296, 1999.

[20] D. Sanchez-Cortes, P. Motlicek, and D. Gatica-Perez. Assessing the impact of language style on emergent leadership perception from ubiquitous audio. In *Proc. of MUM*, Dec. 2012.

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