Noname manuscript No.
(will be inserted by the editor)

Computational Personality Recognition in Social Media

Golnoosh Farnadi · Geetha Sitaraman · Shanu Sushmita · Fabio Celli · Michal Kosinski · David Stillwell · Sergio Davalos · Marie-Francine Moens · Martine De Cock

Received: date / Accepted: date

Abstract A variety of approaches have been recently proposed to automatically infer users' personality from their user generated content in social media. Approaches differ in terms of the machine learning algorithms and the feature sets used, type of utilized footprint, and the social media environment used to collect

This paper or a similar version is not currently under review by a journal or conference, nor will it be submitted to such within the next three months. This paper is void of plagiarism or self-plagiarism as defined in Section 1 of ACM's Policy and Procedures on Plagiarism (`http://www.acm.org/publications/panel/policies/plagiarism_policy`).

Golnoosh Farnadi
Dept. of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
and Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium
E-mail: golnoosh.farnadi@ugent.be

Geetha Sitaraman and Shanu Sushmita
Center for Data Science, University of Washington Tacoma, USA
E-mail: sgeetha@uw.edu and E-mail: sshanu@uw.edu

Fabio Celli
University of Trento, Italy
E-mail: fabio.celli@unitn.it

Michal Kosinski
Stanford University, Stanford, USA
E-mail: michalk@stanford.edu

David Stillwell
The Psychometrics Centre, University of Cambridge, UK
E-mail: ds617@cam.ac.uk

Sergio Davalos
Milgard School of Business, University of Washington, Tacoma, USA
E-mail: sergiod@uw.edu

Marie-Francine Moens
Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium
E-mail: sien.moens@cs.kuleuven.be

Martine De Cock
Center for Data Science, University of Washington Tacoma, USA
E-mail: mdecock@uw.edu

the data. In this paper, we perform a comparative analysis of state-of-the-art computational personality recognition methods on a varied set of social media ground truth data from Facebook, Twitter and YouTube. We answer three questions: (1) Should personality prediction be treated as a multi-label prediction task (i.e., all personality traits of a given user are predicted at once), or should each trait be identified separately? (2) Which predictive features work well across different on-line environments? and (3) What is the decay in accuracy when porting models trained in one social media environment to another?

Keywords Big Five personality · Social media · User generated content · Multivariate regression · Feature analysis

1 Introduction

Research in psychology has suggested that behavior and preferences of individuals can be explained to a great extent by underlying psychological constructs: personality traits [42]. Knowledge of an individual's personality allows us to make predictions about preferences across contexts and environments, and to enhance recommendation systems [33]. Personality can affect the decision making process and has been shown to affect preferences for websites [31], products, brands and services [32], and for content such as movies, TV shows, and books [9].

The most widely accepted model of personality, Big Five or Five Factor Model, embraces five traits [12]: Openness, Conscientiousness, Extroversion, Agreeableness, and Emotional Stability (often conversely referred to as Neuroticism). Further explanations of each trait are summarized in Table 1.

A traditional approach to measure personality requires participants to answer a series of questions (typically, from 20 to 360) evaluating their behavior and preferences (e.g. [28,12]). This approach is time-consuming and impractical, especially in the context of on-line services. On-line users might be unwilling to spend a considerable amount of time filling-in a questionnaire, in order to personalize their search results or product recommendations.

However, it has been recently shown that the digital footprint of users can be used to automatically infer their personality. For example, [32] and [55] showed that automated personality judgments based on Facebook Likes are more accurate than those made by users' friends or even their spouses. Also, [43] showed that similar predictions can be based on language used in social media. A variety of other approaches have been proposed using different prediction mechanisms, feature spaces, and focusing on different on-line environments [11,15,46].

2 Aims of the Study

In this study, we perform a comparative analysis of state-of-the-art computational personality recognition methods on a varied set of social media benchmark datasets collected on Facebook, Twitter and YouTube. Our aim is to address the three following questions.

(1) Should personality prediction be treated as a multi-label prediction task (i.e., all personality traits of a given user are predicted at once), or should each trait be identified separately?

Table 1 Overview of the Big Five Personality Model.

Trait	Description
Openness	Openness is related to imagination, creativity, curiosity, tolerance, political liberalism, and appreciation for culture. People scoring high on Openness like change, appreciate new and unusual ideas, and have a good sense of aesthetics.
Conscientiousness	Conscientiousness measures preference for an organized approach to life in contrast to a spontaneous one. People scoring high on Conscientiousness are more likely to be well organized, reliable, and consistent. They enjoy planning, seek achievements, and pursue long-term goals. Non-conscientious individuals are generally more easy-going, spontaneous, and creative. They tend to be more tolerant and less bound by rules and plans.
Extroversion	Extroversion measures a tendency to seek stimulation in the external world, the company of others, and to express positive emotions. People scoring high on Extroversion tend to be more outgoing, friendly, and socially active. They are usually energetic and talkative; they do not mind being at the center of attention, and make new friends more easily. Introverts are more likely to be solitary or reserved and seek environments characterized by lower levels of external stimulation.
Agreeableness	Agreeableness relates to a focus on maintaining positive social relations, being friendly, compassionate, and cooperative. People scoring high on Agreeableness tend to trust others and adapt to their needs. Disagreeable people are more focused on themselves, less likely to compromise, and may be less gullible. They also tend to be less bound by social expectations and conventions, and more assertive.
Emotional Stability	Emotional Stability, reversely referred to as Neuroticism, measures the tendency to experience mood swings and emotions such as guilt, anger, anxiety, and depression. People scoring low on Emotional Stability (high Neuroticism) are more likely to experience stress and nervousness, while people scoring high on Emotional Stability (low Neuroticism) tend to be calmer and self-confident.

Given the user generated content of each user, the aim is to obtain a set of five estimates (real numbers) representing the Big Five dimensions. We treat this problem as a regression problem by exploring different univariate and multivariate regression techniques. Recently, research has been done on the use of multivariate regression for personality prediction on Facebook [27] and YouTube [15].

In this study, we compare multivariate regression techniques, e.g., multi-target stacking, ensemble of regressor chains, and multi-objective random forests [54], with univariate approaches such as support vector machines and decision trees, as well as with an average baseline algorithm. The average baseline predicts for each data point the mean value across the training data (e.g. if the average openness score of all users in the training data is 2.5, then it predicts that value as the openness score for all users in the test data).

(2) Which predictive features work well across different on-line environments?

We extract a wide variety of linguistic and emotional features from Facebook status updates, tweets and transcripts of vlogs (i.e., video blogs). The underly-

ing rationale for including linguistic and emotional features is that people with different personality traits will express themselves differently and, hence, will use different words (phrases) and emotions (anger, joy).

We assess the strength of the relationship between different predictive features and the personality traits by determining their correlations. We compare the correlation results across different datasets. Finding correlations of text-based features with personality traits has been previously studied (e.g., [43,51,46]). However, to the best of our knowledge, there is no work that compares the results over different benchmark datasets. We select features according to their relationship with personality scores. Motivated by previous research, and the observed correlation between features and personality scores, we include them in our regression models. Our aim is to determine which relationships between features and personality traits are common across various social media platforms.

(3) What is the decay in accuracy when porting models trained in one social media environment to another?

Personality predictions are challenging; unlike demographic data, ground truth (i.e., questionnaire scores) is relatively scarce and measured with a considerable error. Farnadi et al. [16] suggested cross-learning, or developing personality prediction models using a variety of digital environments. The advantage of cross-learning is that training examples from different social media platforms can be combined to increase the accuracy on other test data. Such models could also be applied to environments where training data representative for the deployment domain is not available. In this study, we explore the possibilities of cross-learning for personality prediction by using benchmark datasets from three different environments (i.e., Facebook, YouTube and Twitter).

3 Related Work

In this section we present background material that supports this study. In particular, state-of-the-art efforts related to users' personality predictions, their associated preferences and behavior are provided. In addition, we also describe related work that uses different social media data like Facebook, Twitter and YouTube for the purpose of personality prediction tasks and their analyses.

Personality Prediction, Preference and Behavior: Knowledge about an individuals' personality can allow us to make predictions about preferences across contexts and environments, and enhance recommendation systems [25,41]. Previous work in the field of psychology and human computer interaction (HCI) has highlighted the importance of identifying users' personality traits and their preferences. This can help in building adaptive and personalized systems in order to provide rich and improved user experiences [40]. For instance, in order to understand the online profile creation process, Counts and Stecher [13] conducted a study, and found that free-form profile attributes allow best desired self presentations, and only specific attributes were needed for sufficient self presentation. In a separate study by Lee and Nass [35], interaction effects between user factors, and media factors on feelings of social presence were investigated. It was found that matching synthesized voice personality to the user personality positively affects users' (especially extrovert users') feelings. Such findings can be critical in the design of virtual reality systems and human computer interfaces. In a study

by Saati et al. [50] it was found that extroverts tended to interact faster with the user interface than introverts. The study also suggests that personality data could help designers to select appropriate skin colours for the user interface.

Identifying users' personality is not only useful for commercial purposes, but it can also help in understanding the mental health and high risk factors of online users. For instance, [18] examined the relationship of Social Networking Sites (SNS) and their problematic usage with regard to personality characteristics and depressive symptomatology. The results of this study indicate that problematic SNS usage is significantly and positively related to depression and Neuroticism, while being negatively associated with Agreeableness.

Social Media and Personality: Social media websites provide a unique opportunity for personalized services to capture various aspects of user behavior. Besides users' structured information contained in their profiles, e.g., demographics, users produce large amounts of data about themselves in a variety of ways including textual (e.g., status updates, blog posts, comments) or audiovisual content (e.g., uploaded photos and videos). Many latent variables such as personalities, emotions and moods — which, typically, are not explicitly given by users — can be extracted from user generated content (see e.g. [4,16,20]). Research into automatic personality prediction using social media data is a very nascent area which is gaining increased research attention due to its potential in many computational applications.

Next, we discuss the relevant background material on how different social media data like Facebook, Twitter and YouTube have been used individually by researchers for the purpose of personality prediction tasks and analyses. Note that, in this study, our aim is to perform a comparative analysis of state-of-the-art computational personality recognition methods on a varied set of social media ground truth data from Facebook, Twitter and YouTube.

Facebook Dataset and Personality: In recent years there have been several dedicated research efforts that utilized Facebook data collected as part of the *myPersonality* project (e.g., [23,3,47,16,9,14]). The details of this dataset are described in Section 4.1. In a study by Hagger-Johnson et al. [23], extracted data from the interests and activities sections of Facebook profiles were used to compare general personality and Sensational Interests Questionnaire (SIQ) scores. Sensational interests are interests that are unusually violent such as weapons, martial arts, etc.

Bachrach et al. used the myPersonality Facebook dataset to investigate how users' activity on Facebook relates to their personality. One of the findings was that Neuroticism has a generally significant negative correlation with the number of friends. The results also showed some evidence that Agreeableness is positively correlated with the number of tags. In a study by Farnadi et al. [16] the relation between emotions expressed in Facebook status updates and the users' age, gender and personality were investigated. Several interesting observations were made through this study. For instance, it was found that extrovert and open users are more emotional in their status posts than neurotic users. Another example of research that utilized the myPersonality Facebook data is the study by Cantador et al. [9]. The authors used the Facebook dataset to investigate the relations between personality types and user preferences in multiple entertainment domains, namely movies, TV shows, music, and books. In this paper, we also use the Facebook dataset from the myPersonality project.

Twitter Dataset and Personality: User generated content on Twitter (e.g., tweets) also provides a valuable source of information for inferring users' personality traits. One of the Twitter datasets often used in the literature is collected through the myPersonality project. Among thousands of participants involved in the myPersonality project, only a few hundreds of users posted links to their Twitter accounts, which forms the content of this dataset. This dataset has been used for the task of automatically predicting the personalities of the users, as well as for user behavior analyses [46,26,19]. For instance, Quercia et al. [46] found that extroverts and emotionally stable people are popular as well as influential users on Twitter. It was also observed that popular users are imaginative, while influential people on Twitter are more organized. Golbeck et al. 2011 [19] used profile information from the dataset as features when training machine learning algorithms to predict scores on each of the five personality traits that were predicted within 11% – 18% of their actual value. On the other hand, Hughes et al. 2012 [26] collected a different dataset from Twitter through an advertisement posted on both Twitter and Facebook. The findings of their study revealed a differential relationship between behaviors on Facebook and Twitter. It was also found that there were personality differences between those who have a preference for Facebook or Twitter, suggesting that different people use the same sites for different purposes. The Twitter dataset that we collected for this study (described in Section 4.3) is a new dataset, hence no previous works are based on it.

YouTube Dataset and Personality: Analysis of video content appears to be one of the least studied problems in the domain of computational personality recognition [6]. A recently collected and annotated YouTube dataset (see Section 4.2 for a detailed description) has sparked interest in personality recognition of vloggers (i.e., video bloggers). The task at hand is different from the work on computational personality recognition in the other social media platforms described above, in the sense that the ground truth data does not come from the vloggers themselves, but from other users watching the videos made by the vloggers. In other words, the task being addressed is not recognition of the true personality traits of vloggers, but *predicting how the personality of vloggers is perceived by their viewers.*

For instance, Aran and Gatica-Perez [2] used this data for a comparison between the personality traits extracted from YouTube and in face-to-face meetings. In another study [7], the vlog dataset was used to build personality models trained on the vlogs, and then applied to classify the EAR audio corpus. Their results suggest that while there are inherent differences between the datasets themselves, it does appear that personality is projected in a fundamentally different way between corpora. The YouTube dataset has also been used in the Workshop on Computational Personality Recognition 2014 [10]. The goal of the workshop was to allow participants to compare the performance and quality of different approaches in personality recognition tasks, as well as defining the state-of-the-art. In this paper, we also use this dataset in our experiments.

4 Datasets

Analyses presented in this paper employ three datasets collected from the most popular social media platforms (i.e., Facebook, Twitter and YouTube). All of those datasets are available to other researchers and hence could be used to benchmark

Table 2 (Table on the left) Characteristics of 3731 users in the myPersonality dataset. (Table on the right) Mean and Standard Deviation of Big Five personality scores of 3731 users (range $[1, 5]$).

	Female	Male
# Users	1492	2239
Average age	25	25
Avg Network size	311	309
Avg # Likes	183	184
Avg # Diads	219	227
Avg # Education	2	2
Avg # Status Updates	176	185
Avg # Groups	34	34

Personality	Mean	Std Dev
Extroversion	3.60	.81
Openness	3.90	.66
Agreeableness	3.60	.70
Conscientiousness	3.50	.74
Neuroticism	2.73	.80

new methods and approaches. Besides for their availability, we choose these three datasets for their differences in size, users, and approach of labeling with personality scores to obtain the ground truth data.

Besides the datasets that we use in this study, there are a few golden standard datasets which are publicly available, such as the essay dataset collected by Mairesse et al. [36] and the mobile personality dataset collected by Aharoni et al. [1]. However, these datasets are not social media datasets, thus we do not leverage them in this study.

There are not many golden standard datasets from social media platforms available for the personality prediction task. The main reason is that gathering labeled data is time-consuming and expensive. So far, two approaches have been used to collect personality scores. The first approach requires participation of users to provide self-reported personality via answering questionnaires. This approach has been used to gather labels for the Facebook and Twitter datasets that we use in this study. Another approach is by asking other users for their opinion regarding the personality of a user. Unlike many tasks in natural language processing where labeling data by using human resources is accurate and straightforward, assigning personality scores is a challenging task for non-experts. Using questionnaires to collect perceived personality scores can ease the task, however judging the personality of another person by employing the written or spoken text is a challenging task. even for experts. Collecting personality scores of users via face-to-face interactions or observing each other's behavior is somewhat easier; our YouTube Vlogger dataset is labeled in this way. The rest of this section provides a detailed description of the datasets that we use in this study.

4.1 MyPersonality: Facebook Dataset

MyPersonality [52] was a popular Facebook application introduced in 2007 allowing its users to take a number of psychometric tests, including a standard Five Factor Model questionnaire [21]. Users received feedback on their scores and could opt-in to donate their scores and Facebook profile data to research. Data for over 6 million myPersonality users is available to researchers at: http://mypersonality.org/. It contains scores on more than 20 psychological tests, demographic profiles, and Facebook profile data including status updates, Likes, social networks, views, work and education history and much more.

Table 3 (Table on the left) Characteristics of the 404 users in the YouTube vloggers dataset. (Table on the right) Mean and standard deviation of the perceived Big Five personality scores of the users (range $[1, 7]$).

	Characteristics
# users	# Female - 210 # Male - 194
# AV features	# Audio - 21 # Video - 4
Transcripts	10K unique words 240K word tokens Avg 595 words/transcript

Personality	Mean	Std Dev
Extroversion	4.62	.98
Openness	4.66	.72
Agreeableness	4.68	.88
Conscientiousness	4.50	.77
Emotional Stability	4.77	.80

The sample that is used in this work includes 3731 users who chose English as their default language and has the following information available: age, gender, personality scores, Facebook activities (i.e., counts of Likes; counts of status updates posted by the user; counts of education; counts of diads from the friendship diads table of the user; counts of group memberships for the user; and network size or number of friends of the user); and at least one status update.

Since our goal is to infer the Big Five personality scores for a given user, we identify a user with his or her set of available status updates which are treated as one text per user, his or her demographic features and Facebook activities. The Big Five personality scores for each user are available in the range of $[1, 5]$. Table 2 provides details about this dataset's characteristics.

Note that the sample that we use is not the largest possible sample from the myPersonality data consisting of users with all the mentioned information. We randomly selected a sub-sample which is large enough for analysis (an order of magnitude larger than the YouTube dataset, which we will discuss next), while at the same time small enough to process with the tools that we leverage in this study. Investigating tools for big data analysis is out of the scope of this paper, thus we leave it for our future work.

4.2 YouTube Vlog Dataset

A video blog or video log, usually abbreviated as vlog, is the video form of a blog. Vloggers explicitly show themselves in front of a webcam, talking about a variety of topics including personal issues, politics, movies, books, etc. Figure 1 shows an excerpt from the transcript of a vlog. The YouTube Vlog dataset[1] that we use in this study was collected by Biel et al. in 2011 [6,7], and consists of 404 vlogs. For each vlog, 25 *audio-video features* are available, as well as a raw text *speech transcript* corresponding to the full video duration, the *gender* of the vlogger, and *personality impression scores*. Table 3 provides details about this dataset's characteristics and personality scores mean and standard deviation.

The *personality impressions* consist of Big Five personality scores that were collected using Amazon's Mechanical Turk (MTurk) crowd sourcing platform and the Ten-Item Personality Inventory (TIPI). MTurk annotators watched one-minute slices of each vlog, and rated impressions using a personality questionnaire. The

[1] https://www.idiap.ch/dataset/youtube-personality

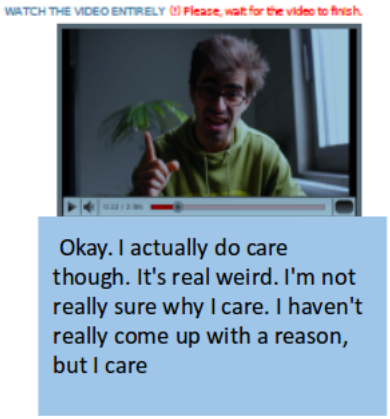

Fig. 1 An example of an excerpt from a vlog transcript

Big Five personality impression scores are available for each user over all the five traits in the range of $[1, 7]$.

The *audio-video features* were automatically extracted from the conversational excerpts of the vlogs and aggregated at the video level. The video features were extracted from the vloggers body activities and include 4 features: the entropy, median, and center of gravity in horizontal and vertical dimensions. The 21 audio features include speaking time, length of the speaking segments, number of speaking turns, voicing rate, ratio of looking while speaking, ratio of looking while not speaking, and multimodal ratio, in addition to mean and standard deviation of speaking energy, pitch, looking time, length of the looking segments, number of looking turns, proximity to the camera, and vertical framing. For more details we refer to [6].

4.3 Twitter Dataset

The Twitter dataset consists of a small set of 102 Twitter users, labeled with gold-standard self-assessed personality types in the range of $[-0.5, 0.5]$. Users have been recruited by means of a Twitter advertising campaign in different languages and their personality types have been assessed with the 10-item personality test (BFI-10) [49], which is available in the selected languages.[2] In addition to personality types, we collected age and gender of the Twitter users, and a set of other metadata about them. The statistics of the data that we collected are reported and described in Table 4.

Since our Twitter dataset is multi-lingual, we first detect English speaking users with a language detector. Table 5 presents the distribution of the detected languages in our Twitter dataset. The sample we use in the remainder of this paper includes the 44 English speaking users. For each user we have the age and gender, in addition to their tweets.

[2] https://www.ocf.berkeley.edu/ johnlab/bfi.htm

Table 4 (Table on the left) Characteristics of the 102 users in the Twitter dataset. (Table on the right) Mean and standard deviation of the self-reported Big Five personality scores of the users (range $[-0.5, 0.5]$).

	Info
#Users	102
#Words	30K tokens
Average tweets per user	19
#Males	47
#Females	55
Average age	27

Personality	Mean	Std Dev
Extroversion	.16	.18
Openness	.10	.24
Agreeableness	.14	.16
Conscientiousness	.11	.17
Emotional Stability	.23	.19

Language	**# users**
English(en)	44
Spanish(es)	36
Dutch(nl)	13
Italian(it)	7
Portuguese(pt)	1
Somali(so)	1
	102

Table 5 Distribution of languages in the Twitter dataset

5 Methodology

5.1 Extracted Features

We extracted a wide variety of linguistic and emotional features from the three datasets that we use in this study. Pychological studies [36] show that there exist links between linguistic features (extracted from text and conversations) and users' personality traits. This finding is demonstrated by the correlations between features such as acoustic parameters, lexical categories, and n-grams on one hand, and the personality classes on the other hand [44]. As a result, it has become increasingly popular to use language in social media for predicting personality. These findings motivate the choice of the following **Linguistic Features** extracted from text that we use in our experiments. In the rest of this section, when we refer to *document*, we mean the combination of all the status updates of a user in the case of the Facebook dataset, the combination of all tweets of a user in the case of the Twitter dataset and the transcript of a vlog for the case of the YouTube dataset.

- **LIWC**: the Linguistic Inquiry and Word Count tool, known as LIWC, is well-known text analysis software which is widely used in psychology studies [44]. Using the LIWC tool, we extracted **81 features** from each document including features related to standard counts (e.g., word count), psychological processes (e.g., the number of anger words such as *hate* and *annoyed* in the document), relativity (e.g., the number of verbs in the future tense), personal concerns (e.g., the number of words that refer to occupation such as *job* and *majors*), and linguistic dimensions (e.g., the number of swear words). For a complete overview of the features, we refer to [53].
- **NRC:** NRC is a lexicon that contains more than 14,000 distinct English words annotated with 8 emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust), and 2 sentiments (negative, positive) [37]. For each document

we counted the number of words in each of the 8 emotion and 2 sentiment categories, resulting in **10 features** per document. The NRC Emotion Lexicon has been used in other works for the task of personality predictions, e.g. [38] and [15]. The underlying rationale for including emotional features (NRC) is that people with different personality traits will express themselves differently and, hence, will use different words (phrases) and emotions (such as anger and joy). A relation between emotions and personality traits has been observed in past research as well [14].

- **MRC**: MRC is a psycholinguistic database[3] which contains psychological and distributional information about words. The MRC database contains 150,837 entries with information about 26 properties (e.g., the number of syllables in the word, the number of letters, etc.), although not all properties are available for every word. Using MRC we generated **14 features** for every document by adding the MRC-scores for each word in the document. Extracted features are: number of letters in the word (Nlet), number of phonemes in the word (Nphon), number of syllables in the word (Nsyl), Kucera and Francis written frequency (KF freq), Kucera and Francis number of categories (KF ncats), Kucera and Francis number of samples (KF nsamp), Thorndike-Lorge frequency (TL freq), Brown verbal frequency (BROWN freq), Familiarity (Fam), concreteness (Conc), imagery (Imag), mean Colerado Meaningfulness (Meanc), mean Pavio Meaningfulness (Meanp), and age of acquisition (Aoa). MRC features used in previous studies such as [17] showed that there is a significant correlation between *Extroversion* and concreteness features, as well as between *Conscientiousness* and words expressing insight, longer words (Nphon, Nlet, Nsyl and Sixltr), and words that are acquired late by children (Aoa) in the MRC database.
- **SentiStrength:** SentiStrength[4] assigns to each text a positive, negative and neutral sentiment score on a scale of 1 (no sentiment) to 5 (very strong sentiment). Texts may be simultaneously positive, negative and neutral. We used SentiStrength to compute 2 sentiment scores (**2 features**) for every document. There are different ways to get the output from SentiStrength. For this study we chose "dual", in which for each given text we get two values corresponding to negative and positive sentiment, and the neutral score can be calculated by summing these two numbers. We disregarded the neutral score in our study. Many studies have successfully exploited emotion and sentiment features in personality prediction tasks such as [10, 15].
- **SPLICE:** We used SPLICE[5] (Structured Programming for Linguistic Cue Extraction) to extract **66 linguistic features**, including cues that relate to the positive or negative self evaluation of the speaker (e.g., *I'm able*, *don't know*), complexity and readability scores. SPLICE features have also been used in a number of psychological studies and personality prediction tasks including [15].

For the Facebook dataset, we extracted features from one textual document file per user. The complete list of the extracted features from the Facebook dataset includes the demographic features, i.e., age and gender, the Facebook activity

[3] http://www.psych.rl.ac.uk/User_Manual_v1_0.html
[4] http://sentistrength.wlv.ac.uk
[5] http://splice.cmi.arizona.edu

features as explained in Table 2, such as the number of likes and status updates, and the linguistic features except for the NRC features. For the YouTube dataset, in addition to the given audio/video and gender features, for each vlogger we extracted all the linguistic features from the vlogs' transcripts. And finally, similar to the Facebook dataset, for the Twitter dataset, we have the age and gender of users and we extracted all the linguistic features, except for the NRC features, from the users' tweets.

The NRC features are not extracted from the Facebook statuses and tweets. Emotion is a momental feeling with respect to an object, person, event, or situation. As a consequence, people express a variety of different emotions over a period of time. Since we combine all status updates or tweets of a user to extract linguistic features, extracting NRC features without considering the context is irrelevant.

In this study, we extract dictionary-based linguistic features, also known as closed-vocabulary approaches, to compare the predictive ability of features across different social media platforms. Open-vocabulary linguistic features for the task of personality prediction have been studied as well in previous work, with promising results such as in [51]. Examples of open-vocabulary features are n-grams, clustered groups of semantically related words (e.g., latent Dirichlet allocation (LDA) topics), and differential language analysis (i.e., DLA).

Unlike open-vocabulary approaches, the quality and processing time of the features extracted by the dictionary-based approaches do not depend on the size of the data. However, one limitation of using dictionary-based linguistic features for the task of personality prediction in social media is the dynamic and noisy structure of these platforms. Users in social media tend to use informal language which contains language errors, misspelled words and newly defined terms and phrases. Thus, improving the performance of the dictionary-based approaches on user generated texts in social media is an open path to explore.

5.2 Regression Approaches

Regression is the task of predicting a continuous, real valued output from a set of predictors. As the name implies, univariate regression refers to estimating a regression model with one dependent variable (one outcome), while multivariate regression refers to building a regression model with more than one dependent variable (several outcomes). The results in Table 6 indicate a clear correlation among different personality trait scores in the YouTube, Facebook and Twitter datasets. The dependency among different personality scores makes personality score prediction a good candidate for multivariate regression, where the dependencies between the target variables are taken into account to make a combined prediction.

Formally, univariate/multivariate regression addresses this problem: let $\mathcal{F}$ be the input space consisting of vectors with values for m features, $f_1, f_2, ..., f_m$, and let $\mathcal{T}$ be the output space consisting of vectors with values for n target variables $t_1, t_2, ..., t_n$. The goal of a multivariate regression algorithm is to learn a model $\mathbf{M} : \mathcal{F} \rightarrow \mathcal{T}$ that minimizes the prediction error over a train set.

In this study, $n = 5$ (where t_1 is *Extroversion*, t_2 is *Agreeableness*, t_3 is *Conscientiousness*, t_4 is *Emotional Stability/Neuroticm* and t_5 is *Openness*). Using this

Table 6 Pearson product-moment correlation results among personality scores on five traits: *Extroversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Emotional Stability (Ems) vs. Neuroticism (Neu)*, and *Openness (Open)* on the Facebook dataset, YouTube vloggers dataset and Twitter dataset. Significant correlations ($p < .05$) among the personality scores are indicated in bold.

Facebook					
	Extr	Agr	Cons	Neu	Open
Extr	1.00				
Agr	**.17**	1.00			
Cons	**.16**	**.18**	1.00		
Neu	**-.32**	**-.33**	**-.28**	1.00	
Open	**.14**	**.04**	-.01	**-.05**	1.00
YouTube					
	Extr	Agr	Cons	Ems	Open
Extr	1.00				
Agr	.02	1.00			
Cons	-.03	**.38**	1.00		
Ems	.06	**.69**	**.54**	1.00	
Open	**.56**	**.29**	**.26**	**.30**	1.00
Twitter					
	Extr	Agr	Cons	Ems	Open
Extr	1.00				
Agr	**0.27**	1.00			
Cons	0.01	0.1	1.00		
Ems	**0.46**	**0.34**	0.15	1.00	
Open	-0.05	-0.06	0.1	0.05	1.00

formulation, the univariate and multivariate regression algorithms that we use in this paper are [54]:

1. **Single-Target (ST)**: In ST, for each target variable t_i, a single model $M_i : \mathcal{F} \rightarrow \mathcal{T}_i$ is trained that maps a vector from the input space $\mathcal{F}$ to a value in $\mathcal{T}_i$, which is the range of variable t_i. The results of the desired multi-target model **M** are comprised of the outcomes of the single-target models.
2. **Multi-Target Stacking (MTS)**: MTS consists of two steps. In the first step, n single-target models are used as in ST, however, MTS includes an additional step where the input space for each target variable is expanded by the predicted results of the other target variables ($n-1$ predicted values) from step one. Let $t_1^{'}, t_2^{'}, ..., t_n^{'}$ be the prediction results from the first step, then, for example, the input space for t_1 in step two is $[f_1, f_2, ..., f_m, t_2^{'}, t_3^{'}, ..., t_n^{'}]$.
3. **Multi-Target Stacking Corrected (MTSC)**: In MTSC, an internal cross-validation sampling technique is used to avoid over-estimation of the training set. In MTSC, by using k-fold sampling, the prediction results of $\frac{k-1}{k}\%$ of the whole training set are used to expand the input space in the second step as in MTS. In this study we use $k = 10$.
4. **Ensemble of Regressor Chains (ERC)**: The idea behind ERC is chaining single-target regression models. By choosing an order for the target variables (e.g., $O = (t_1, t_2, ..., t_n)$), the learning model for each target variable t_j relies on the prediction results of all target variables t_i which appear before t_j in the list. For the first target variable, a single-target regression model as in ST predicts the value, then the input space for the next target variable is extended with the prediction results of the previous one and so on. Since in this model

the order of the chosen chain affects the results, the average prediction result of r different chains (in our study we choose $r = 10$, as is typically done) for each target variable is used as the final prediction result.

5. **Ensemble of Regressor Chains Corrected ($ERCC$)**: The difference between ERC and $ERCC$ is similar to that between MTS and $MTSC$, i.e., the use of k-fold sampling to increase the reliability of the predictions based on the training set. In this study we use $k = 10$.
6. **Multi-objective random forest (MORF)**: $MORF$ is a random forest ensemble technique of multi-objective decision trees ($MODT$). Each $MODT$ is a multi-target regression model that predicts multiple target variables at once. $MODT$ models are instantiations of predictive clustering trees ($PCTs$) that are used for multi-objective prediction [8]. The PCTs algorithm and standard decision trees differ in the way they treat the variance and the prototype functions. In $PCTs$, the variance and the prototype functions are treated as parameters, and they are instantiated towards a given prediction task for computing the leaf labels. For multi-objective regression trees, the variance is computed as the sum of the variances of the target variables (t_i). That is, $Var(E) = \sum_{i=1}^{n} Var(t_i)$, where E is a set of training examples, and each leaf's prototype is the vector mean of the target vectors of its training examples. Multi-objective random forests ($MORF$) have shown better predictive performance than their counter ensemble methods like bagging for $MODT$ [30].

Note that ST does not leverage the prediction result for one personality trait to make a prediction for another, while all other algorithms (MTS, $MTSC$, ERC, $ERCC$ and $MORF$) do in one way or another. To get the results for ST, MTS, $MTSC$, ERC, $ERCC$ and $MORF$ we used the implementation of these algorithms in Mulan[6]. The base learner of these algorithms in Mulan (except $MORF$, as explained above) is the Weka decision tree algorithm. For further information we refer to [54]. For the ERC and $ERCC$ models we choose 10 randomly selected chains and for $MORF$ we use an ensemble size of 100 trees. For the rest of the parameters we use the suggestions in [30].

We also use the R software environment [48] to implement ST and MTS with a support vector machine regressor with radial kernel as a base learner. In the remainder of this paper we mention the base learner in parentheses after the approach name to make it clear which base learner is used, for example, MTS (SVM) refers to the Multi-Target Stacking approach with a support vector machine regressor as a base learner. In the case of SVM, we tried different kernels, namely radial, linear and polynomial, and tuned the parameters based on the training set. Since we obtained the best results with a radial kernel, all results presented throughout this paper that are based on a SVM as base leaner rely on a radial kernel.

5.3 Evaluation Approaches

We evaluate the results based on *Root Mean Squared Error (RMSE)* and *Coefficient of Determination (R^2)*. RMSE measures the difference between the pre-

[6] http://mulan.sourceforge.net/

dicted values by a model and the observed values. RMSE ranges from 0 to ∞ where lower values signify better models. RMSE can be described by the following formula:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_{obs}^{t} - y_{pred}^{t})^2}{n}} \quad (1)$$

where y_{obs}^{t} and y_{pred}^{t} are the observed and predicted scores for instance t (where $t = 1 \ldots n$) and n is the sample size.

R^2 is the ratio of the model's absolute error and the baseline mean predicted scores. It is expressed as:

$$R^2 = 100 \times \left(1 - \frac{\sum_{t=1}^{n}(y_{obs}^{t} - y_{pred}^{t})^2}{\sum_{t=1}^{n}(y_{obs}^{t} - \hat{y}_{obs})^2}\right) \quad (2)$$

where y_{obs}^{t} and $\hat{y}_{obs}$ are respectively the observed scores and their mean, and y_{pred}^{t} are the predicted scores by the model. R^2 measures the relative improvement of the mean squared error using the automatic predictor compared to the average baseline. Positive values indicate that the model accounts for a greater proportion of the variance in the data thus outperforming the constant average baseline. Negative values indicate that variation in the data accounted for by the model is worse than the baseline score, thus not outperforming the baseline.

6 Experimental Results

In this section, we present the details of the experiments and the results of personality prediction using our three social media datasets.

6.1 Correlation Results

We perform pair-wise correlation analysis between the extracted features and personality scores for all three datasets. In particular we use the non-parametric Spearman rank correlation to compute the correlations in the YouTube and Twitter datasets due to the non-normal and highly skewed nature of the distribution of individual features. For the Facebook dataset we use the parametric Pearson correlation when reporting the correlations. For computing Spearman and Pearson correlations between the features and the five personality scores, we use the R software environment [48].

Table 7 contains a summary of the most important correlation results across all three social media datasets.All the presented correlation results are significant with $p < 0.05$.[7] The demographic features *age* and *gender* have a significant correlation with personality scores across all three datasets. Following a commonly adopted encoding approach, in our experiments, gender equal to 1 indicates female users and 0 indicates male. In fact correlation with the gender feature is simply

[7] We compute the correlation among all features and personality traits and find the significant correlated features. The full list of features and their correlation scores can be downloaded from the supplementary materials of this manuscript.

Table 7 Common significantly ($p < .05$) correlated features with personality traits. The personality traits: *Extroversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Emotional Stability (Ems) vs. Neuroticism (Neu)*, and *Openness (Open)*, across Facebook, YouTube and Twitter datasets. The significant features after Bonferroni-correction (with $p < .01$) are typeset in bold.

Feature	Trait	Facebook	YouTube	Twitter
Demographics				
Gender	Agr	**0.06**	**-0.24**	-0.18
Age	Ems/Neu	**0.04**	-	0.32
Age	Agr	0.04	-	0.41
LIWC				
WC (word count)	Agr	0.02	-0.11	0.31
negate	Cons	-0.03	-0.22	-0.42
health	Ems/Neu	-0.04	-0.11	0.31
assent	Extr	0.03	**0.17**	0.33
motion	Open	-0.02	0.11	-0.31
leisure	Ems/Neu	0.04	0.12	**0.43**
MRC				
AOA	Cons	0.04	**0.16**	0.33
NLET	Agr	0.05	-0.11	0.31
SPLICE				
num Adjectives	Agr	**0.04**	-0.13	0.30
SWN Positivity	Agr	0.05	**0.19**	0.32
SWN Negativity	Agr	-0.02	**-0.20**	0.37

a comparison of the means of personality scores for men and women. An appropriate approach to calculate this association is point-biserial correlation which is mathematically equivalent to the Pearson correlation by using 0/1 values. Thus, we use the Pearson correlation for finding the relations among personality traits and gender for all three datasets.

There is a positive relation (0.06) between *gender* and the Agreeableness personality trait on Facebook. However, the relation is negative (-0.24) in case of the YouTube dataset and Twitter dataset (-0.18). This means that for female Facebook users, the mean personality score for Agreeableness will be higher than men, but lower in case of YouTube and Twitter users. In addition, *age* has a similar correlation (0.04) with the Emotional stable and Agreeableness personality scores.

In case of linguistic inquiry and word count (LIWC) features, six features were found to be common and significantly correlated across the three datasets. Similar to demographic features, these LIWC features exhibit different relations depending on the dataset type. For example, the word count (WC) shows a positive relation with the Agreeableness personality score in the Facebook (0.02) and Twitter (0.31) datasets, but is negatively related in the YouTube dataset (-0.11).

There were only two features – age of acquisition rating (AOA), and number of letters ($NLET$) from the MRC psycholinguistic database that are common and have a significant correlation with the personality scores across the three datasets. Both features show a positive relation with Conscientiousness and Agreeableness personality scores for the Facebook (0.04 and 0.05) and Twitter (0.33 and 0.31) datasets. But, a negative correlation is found between the number of letters ($NLET$) and the Agreeableness personality score in the YouTube dataset.

Finally, four features from SPLICE were found to be highly correlated and common among the three datasets. Interestingly, all these features were only cor-

related to the Agreeableness personality score. The relation was positive in the Facebook and Twitter datasets, but mostly negative in the YouTube dataset.

Furthermore, to avoid type 1 error of multiple testings, we apply Bonferroni correction with ($p < .01$). To have a fair comparison among the datasets, we only consider the common non zero features, thus we identify the correlations among 161 common features and 5 personality traits between three datasets. By adjusting the p-values, the number of significant correlations among features and personality traits are decreased. For the case of the Twitter dataset the number of significant features reduces from 51 significant correlated features to 11, for the case of the YouTube dataset the number of significant correlated features drops from 231 to 141, and finally for the case of the Facebook dataset the number of significant correlated features cuts down from 240 to 164 significant correlations.

The relation between the sample size and number of correlated features is addressed in [51] for the personality prediction in Facebook. Similarly, we discover a direct relation between the population size and number of correlated features, i.e. we find 11 significant correlations for the case of the Twitter dataset with only 44 examples, 140 significant correlations for the case of YouTube with 404 samples and 164 correlations for the Facebook dataset with 3731 users. Since the population size affects the number of features, by adjusting the correlations, we do not find any common significant correlated features among all three datasets.

Overall, two key observations can be made from the results in the correlation Table 7. First, not all features are common and significantly correlated to the personality scores. For instance, among the 81 LIWC features, only six features were found to be significantly correlated ($p < .05$) and common in all three datasets. Second, features can have a different relation with the personality score depending on the dataset. In one dataset, a feature can be positively related to a personality score (e.g., gender for Agreeableness in Facebook), while the same feature may have a negative correlation in a different dataset (e.g., gender for Agreeableness in YouTube). This suggests that it may not be possible to generalize the correlation between features and personality traits, as this may vary depending on the social media platform.

6.2 Regression Models

In this section, by using the univariate and multivariate regression formulations that we described in section 5.2, we explore different approaches to computational personality recognition of social media users. We predict personality on a continuous scale which is common in psychology studies. While we predict the *perceived* personality scores from spoken text (transcripts from video) in the YouTube vloggers dataset, we predict the *self-reported* personality scores from written text as status updates and tweets from the Facebook and the Twitter datasets, respectively.

The experimental results using feature selection are presented in Section 6.3 and then results of applying different univariate and multivariate regression formulations are presented in Section 6.4. All results are based on 10-fold cross-validation, where folds are randomly sampled from the data.

Throughout this section, we use letter codes for different regressors as described in Table 8.

Table 8 Regressors and the corresponding letter codes.

Regressor	Base learner	Letter Code
Univariate Regressors		
Single-Target	Decision tree	ST (DT)
Single-Target	Support vector machine	ST (SVM)
Multivariate Regressors		
Multi-Target Stacking	Decision tree	MTS (DT)
Multi-Target Stacking	Support vector machine	MTS (SVM)
Multi-Target Stacking Corrected	Decision tree	MTSC (DT)
Ensemble of Regressor Chains	Decision tree	ERC (DT)
Ensemble of Regressor Chains Corrected	Decision tree	ERCC (DT)
Multi-objective random forest	Multi object decision tree	MORF

6.3 Experiments Using Feature Selection

Previous studies with regard to personality prediction suggest that feature selection can improve the accuracy of learning algorithms [16]. Feature subset selection is the process of identifying relevant features and removing irrelevant and redundant features before training of the model. It has been shown that feature subset selection enhances the performance of learning algorithms by reducing the hypothesis search space and/or reducing the storage or processing requirement [24].

The main focus of our study w.r.t. feature selection and feature correlation analysis in the paper is on understanding and assessing the impact of individual input features on personality prediction. Our goal is to identify features that are most predictive and relevant to the target variable. We have not measured the correlation among input features themselves. While we acknowledge that that would be interesting as it might lead to regressors with higher accuracy and/or a reduced feature space, we consider that beyond the scope of this paper.

Our incentive for performing feature selection based on correlation analysis is that it is a so-called filter based approach. Unlike wrapper or embedded feature selection approaches [22], filter based feature selection does not depend on the underlying learner, therefore our feature analysis results are general and not tied to a specific learner. Finally, we acknowledge that there are many feature construction methods for dimensionality reduction such as basic linear transforms of the input variables (e.g., PCA [29]) that can improve the performance of the learner, but as stated above, we consider this to be beyond the scope of this paper.

We perform experiments by selecting different feature sets. We first grouped features based on their categories and then the relevant subset of features for each category is identified by conducting correlation analysis as explained in Section 6.1. Hence, to select features from each category, we choose the significantly correlated features with a trait with $p < 0.05$. Next, for each feature category, we perform personality score prediction based on the selected features, using single-target regression with SVM as the base learner. All results presented in Table 9 are averaged over 10-fold cross-validation. In every fold, the correlated features are calculated based only on the training examples, hence the correlated features may differ from one fold to another. The results are specific to each social media platform.

In the case of Facebook, we leverage six feature sets in addition to their corresponding correlated feature sets. By "correlated feature set" we mean the subset

Table 9 RMSE Comparison of three datasets including 3731 Facebook users, 404 YouTube vloggers, and 44 Twitter users by applying all features and correlated features under each feature set category. For each feature set category, using the correlated features in a model is shown with ✓while a model which uses all features is marked with ✗. The personality traits are *Extroversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Emotional Stability (Ems) vs. Neuroticism (Neu)*, and *Openness (Open)*. All results are based on 10-fold cross-validation using SVM (radial kernel). In each column, significant differences ($p < .05$) with respect to the baseline are denoted by a $*$ sign, and the lowest RMSEs are typeset in bold. The average baseline is shown with Avg.

Facebook						
Feature set	Correlated	Extr RMSE	Agr RMSE	Cons RMSE	Neu RMSE	Open RMSE
Avg		.807	.699	.735	.786	.661
Activity & Demographics	✗	**.784**	.702	.721	**.768**	.663
Activity & Demographics	✓	.785	.702	.721	.768	.664
LIWC	✗	.803	.693	.723	.779	.652
LIWC	✓	.806	.693	.725	.782	.657
SentiStrength	✗	.807	.697	.734	.786	.664
SentiStrength	✓	.810	.703	.737	.787	.660
MRC	✗	.811	.700	.730	.787	.663
MRC	✓	.809	.699	.729	.785	.661
SPLICE	✗	.807	.699	.730	.785	.664
SPLICE	✓	.810	.701	.736	.788	.665
All	✗	.791	.695	**.717**	.773	**.651**
All	✓	.786	**.692**	.719	.770	.653
YouTube						
		Extr RMSE	Agr RMSE	Cons RMSE	Ems RMSE	Open RMSE
Avg		.980	.880	.773	.780	.719
Gender, Audio & Video	✗	**.842***	.892	.759	.787	.706
Gender, Audio & Video	✓	.868*	.882	.752	.824	.704
LIWC	✗	.930	.781*	**.683***	.753	.710
LIWC	✓	.933	.775*	.695*	.752	.716
NRC	✗	.984	.814*	.757	.767	.712
NRC	✓	1.00	.816*	.774	.774	.712
SentiStrength	✗	.987	.805*	.758	**.741***	.710
SentiStrength	✓	.987	.815*	.774	.746	.716
MRC	✗	.969	.900	.743*	.790	.721
MRC	✓	.975	.920	.746*	.793	.725
SPLICE	✗	.979	.882	.772	.779	.717
SPLICE	✓	.971	.882	.773	.794	.718
All	✗	.979	.882	.773	.780	.717
All	✓	.867*	**.773***	.708*	.742*	**.700**
Twitter						
		Extr RMSE	Agr RMSE	Cons RMSE	Ems RMSE	Open RMSE
Avg		.179	.159	.175	.198	.236
Demographics	✗	.187	.161	.203	.213	.211
Demographics	✓	.213	**.149**	.203	.195	**.202**
LIWC	✗	.181	.160	.175	.208	.253
LIWC	✓	.181	.160	.175	.288	.253
SentiStrength	✗	.184	.156	.174	.193	.256
SentiStrength	✓	.180	.163	**.132**	.194	.235
MRC	✗	.180	.164	.170	.192	.236
MRC	✓	.194	.178	.167	.189	.270
SPLICE	✗	.185	.163	.183	.188	.240
SPLICE	✓	**.173**	.159	.247	.215	.252
All	✗	.181	.165	.183	**.179**	.226
All	✓	.197	.162	.184	.204	.230

of features that was found to be correlated with the personality trait at hand. Results which are presented in Table 9 indicate that Facebook activities and demographics of a user are better predictors in learning the personality of a user compared to their user generated texts, i.e., extracted features from their status updates. For the case of predicting scores for *Extroversion* and *Neuroticism*, using only this feature set is enough to get the lowest RMSE score. However, for the traits *Agreeableness*, *Conscientiousness* and *Openness*, in addition to this feature category, textual features from the combined status updates improve the performance and lead to the lowest RMSE. Among the five different feature sets (i.e., except for the combination of all features as one feature set *All*) that we extracted for this dataset, users' activities and demographics in addition to LIWC features produce the lowest RMSE for predicting personality scores for all five traits.

In the case of the YouTube vloggers, we analyze seven feature sets and their corresponding correlated ones. The audio and video features extracted from the videos, which reflect the actual behavior of the users, are better predictors compared to the linguistic features for predicting the score of *Extroversion*. However, for other traits, the lowest RMSEs are obtained by leveraging the content of the videos by using the linguistic features extracted from the transcripts. For the case of *Agreeableness* and *Openness*, using the combination of linguistic features and audio and video features in the learning process, resulted in the lowest RMSE. And finally, the models that use LIWC features for *Conscientiousness* score prediction and SentiStrength features for inferring the *Emotional Stability* trait show results with the lowest RMSE score. Overall, for the YouTube dataset using only LIWC features produces better prediction results compared to other feature sets.

For the case of the Twitter dataset, we use six feature sets in addition to their corresponding correlated feature sets. It is interesting that from the demographic features, using only age for inferring the *Agreeableness* score and only gender for predicting the *Openness* score outperform the average baseline while for the case of *Emotional Stability*, using the combination of all feature groups as one feature space led to the best performing model which also outperforms the average baseline. Textual features extracted from the tweets, in particular SPLICE features for the case of *Extroversion* and SentiStrength for the case of *Conscientiousness*, reduce the error and outperform the average baseline. For this dataset, due to the small size of the training set, the results obtained using various feature sets are very similar and choosing one feature set that outperforms other feature sets for all traits is not possible.

Overall, for all the traits in all three data sets, we find at least one feature set which outperforms the average baseline. Note that the feature selection approach only considers the significant correlated features. For feature category and and traits combinations for which no significant correlated features were found, we report the same value as in the case that all features in the feature set are used. From the results in Table 9, we can conclude that selecting features using correlation analysis mostly has little or no improvement compared to using the complete feature set.

6.4 Experiments Using Univariate and Multivariate Regression Approaches

Following the formulations of multiple regression approaches in Section 5.2, the formal definition of regression learners for each dataset is presented as follows. Let $\mathcal{F}$ be the input space consisting of feature vectors. The extracted features for each dataset are different as described in Section 5.1. The Facebook feature space $\mathcal{F}^{FB}$ has 171 features, $f_1^{FB}, f_2^{FB}, ..., f_{171}^{FB}$, the YouTube feature space $\mathcal{F}^{YT}$ has 199 features, $f_1^{YT}, f_2^{YT}, ..., f_{199}^{YT}$, and finally, the Twitter feature space $\mathcal{F}^{TW}$ has 165 features, $f_1^{TW}, f_2^{TW}, ..., f_{165}^{TW}$.

Let $\mathcal{T}$ be the output space, containing vectors with values for 5 target variables: t_1 (*Extroversion*), t_2 (*Agreeableness*), t_3 (*Conscientiousness*), t_4 (*Neuroticm or Emotional Stability*) and t_5 (*Openness*). The goal of a multivariate regression algorithm is to learn a model $\mathbf{M} : \mathcal{F} \rightarrow \mathcal{T}$ that minimizes the prediction error $RMSE$ over a test set. The goal of a univariate regression algorithm is to learn five models $M_1 : \mathcal{F} \rightarrow \mathcal{T}_1$ ($Extroversion$), $M_2 : \mathcal{F} \rightarrow \mathcal{T}_2$ ($Agreeableness$), $M_3 : \mathcal{F} \rightarrow \mathcal{T}_3$ ($Conscientiousness$), $M_4 : \mathcal{F} \rightarrow \mathcal{T}_4$ ($Neuroticm/EmotionalStability$), and $M_5 : \mathcal{F} \rightarrow \mathcal{T}_5$ ($Openness$) that minimize the prediction error RMSE over a test set, with $\mathcal{T}_i$ the range of variable t_i (for $i = 1 \dots 5$).

Some initial research has been done on the use of multivariate regression for personality prediction on Facebook [3,27], YouTube [15] and Sina Microblog data [5]. In the current section we investigate whether the promising trend of good results can be extended to our Facebook, YouTube and Twitter datasets. To compare the performance of different regressor approaches, we apply the same set of approaches on all three datasets. We aim to identify which approach is a better predictor for the task of personality prediction regardless of the dataset. The results of all the experiments are summarized in Table 10. All results are averaged over a 10-fold cross-validation, and to measure significant differences in prediction errors between the learned models and the baseline, we conducted two-tailed paired t-tests for the RMSE, and two-tailed single t-tests for R^2 at the $p < .05$ level.

We use two base learners in our experiments, namely a decision tree algorithm and SVM algorithm. By using the whole feature space, univariate regressor ST (DT) always outperforms ST (SVM); similarly multivariate MTS (DT) accomplishes significantly better results compared to MTS (SVM). Although in many studies SVM has been used successfully for inferring personality traits as a classifier or a regressor approach such as [45,16,39], the results presented in Table 10, which are based on three different social media datasets, indicate that the decision tree algorithm is a better predictor approach for this task.

Moreover, it can be seen from the results in Table 10 that all five algorithms (i.e., ST (DT), MTS (DT), $MTSC$ (DT), ERC (DT) and $ERCC$ (DT)) which use the decision tree algorithm as base learner outperform (i.e., have a lower prediction error than) the average baseline model for all five personality traits. In addition, positive values for R^2 are also observed for all the algorithms which further indicates better performance than the average baseline model ($0\% \leq R^2 \leq 33\%$).

An interesting observation is that multivariate regression approaches (i.e., MTS (DT), $MTSC$ (DT), ERC (DT) and $ERCC$ (DT)) not always outperform the univariate approach i.e., ST (DT), but most of the times they give better results. However, the differences between univariate and multivariate regressors

Table 10 Root mean square error ($RMSE$) and Coefficient of determination (R^2) results for personality trait prediction using *univariate* and *multivariate* regression algorithms on all 3 datasets. The personality traits are *Extroversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Emotional Stability (Ems) vs. Neuroticism (Neu)*, and *Openness (Open)*. All results are averaged over a 10-fold cross-validation. In each column, the lowest error and highest determination are typeset in bold. Significant differences with respect to the baseline ($p < .05$) are marked using $*$. The average baseline is shown with *Avg*.

Facebook										
	Extr		Agr		Cons		Ems		Open	
Approach	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2
Avg	.807		.699		.735		.786		.661	
Univariate/Multivariate Regressions using Decision Tree										
ST (DT)	.777	7.30	.691	2.28	**.713**	**5.90**	.765	5.27	**.649**	**3.60**
MTS (DT)	.782	6.10	.698	0.29	.717	4.84	.772	3.43	.650	3.30
MTSC (DT)	.777	7.30	**.690**	**2.56**	.714	5.63	**.763**	**5.77**	**.649**	**3.60**
ERC (DT)	**.776**	**7.54**	**.690**	**2.56**	**.713**	**5.90**	.766	5.12	**.649**	**3.60**
ERCC (DT)	**.776**	**7.54**	**.690**	**2.67**	**.713**	**5.90**	**.763**	**5.77**	**.649**	**3.60**
MORF	.787	4.90	.693	1.71	.720	4.04	.774	3.03	.653	2.41
Univariate/Multivariate Regressions using SVM										
ST (SVM)	.791	3.93	.695	1.14	.717	4.84	.773	3.28	.651	3.00
MTS (SVM)	.802	1.24	.695	1.14	.718	4.57	.789	-.76	.651	3.00
YouTube										
	Extr		Agr		Cons		Ems		Open	
Approach	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2
Avg	.980		.880		.773		.780		.719	
Univariate/Multivariate Regressions using Decision Tree										
ST (DT)	.858*	23.35*	.724*	32.31*	.692*	19.86*	**.696***	**20.38***	.695	6.56
MTS (DT)	.862*	22.63*	.722*	32.69*	.696*	18.93*	.711*	16.91*	**.690**	**7.90**
MTSC (DT)	**.850***	**24.80***	**.720***	**33.06***	**.690***	**20.32***	.700*	19.46*	**.690**	**7.90**
ERC (DT)	**.850***	24.80*	.740*	29.29*	.700*	17.35*	.700*	19.46*	**.690**	**7.9**
ERCC (DT)	.853*	24.23*	.721*	32.87*	**.690***	**20.32***	.697*	20.15*	.693	7.10
MORF	.908	14.15	.771*	23.24*	.699*	18.23*	.719*	15.0*	.703	4.40
Univariate/Multivariate Regressions using SVM										
ST (SVM)	.979	.204	.882	-.45	.773	0	.780	0	.717	.56
MTS (SVM)	.987	-1.43	.896	-3.67	.745	7.11	.786	-1.54	.724	-1.4
Twitter										
	Extr		Agr		Cons		Ems		Open	
Approach	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2
Avg	.179		.159		.175		.198		.236	
Univariate/Multivariate Regressions using Decision Tree										
ST (DT)	**.173**	**6.59**	.152	8.61	.165	11.10	.187	10.80	**.214**	**17.78**
MTS (DT)	.174	5.51	.151	9.81	.165	11.10	.188	9.85	.216	16.23
MTSC (DT)	.174	5.51	.152	8.61	**.164**	**12.18**	.187	10.80	**.214**	**17.78**
ERC (DT)	.174	5.51	.152	8.61	.165	11.10	.187	10.80	**.214**	**17.78**
ERCC (DT)	**.173**	**6.59**	.153	7.40	**.164**	**12.18**	.187	10.80	.219	13.89
MORF	.180	-1.12	**.150**	**11.00**	.170	5.63	.180	17.36	.220	13.10
Univariate/Multivariate Regressions using SVM										
ST (SVM)	.181	-2.25	.165	-7.69	.183	-9.35	.179	18.27	.226	8.29
MTS (SVM)	.181	-2.25	.162	-3.81	.175	0	**.176**	**20.99**	.234	1.69

are not significant. Overall, $ERCC$ (DT) and $MTSC$ (DT) outperform the other approaches across all three different datasets for all five personality predictions.

Although feature selection as suggested in many studies such as [16] can generate promising results for the task of personality prediction, using the full feature space for the results presented in Table 10 indicate that feature selection as we use in this study (Table 9) barely yields any advantage. Overall, $ERCC$ (DT) for all traits in the Facebook dataset, $MTSC$ (DT) for YouTube, and both $ERCC$ (DT) and $MTSC$ (DT) outperform all other approaches in predicting the personality traits and yield a lower RMSE score compared to the average base line.

Finally, while *Agreeableness* followed by *Extroversion* are the easiest personality traits of YouTube vloggers to predict using the observers' score as ground truth, *Extroversion* followed by *Conscientiousness* are the best performing traits using the self-reported personality models of Facebook, and similarly *Openness* followed by *Conscientiousness* are the easiest trait to predict for self-reported personality of Twitter users.

7 Cross-media Learning

In this section, we explore cross-media learning by utilizing the available golden-standard datasets to train models in different platforms when little or no training data is available.

Table 11 Overview of the range of the personality trait scores among the datasets Facebook, Twitter and YouTube. Personality traits are *Extroversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Emotional Stability (Ems) vs. Neuroticism (Neu)*, and *Openness (Open)*.

Trait	**Max score**	**Min score**	**Questionnaire type/range**
Facebook			The Big Five inventory questionnaire
Extr	5	1	[1, 5]
Open	5	1	[1, 5]
Cons	5	1	[1, 5]
Agr	5	1	[1, 5]
Neu	5	1	[1, 5]
Twitter			10-item Personality Test (BFI-10)
Extr	0.5	-0.3	[-0.5, 0.5]
Open	0.5	-0.2	[-0.5, 0.5]
Cons	0.5	-0.2	[-0.5, 0.5]
Agr	0.5	-0.3	[-0.5, 0.5]
Ems	0.5	-0.3	[-0.5, 0.5]
YouTube			Ten-Item Personality Inventory (TIPI)
Extr	6.6	2	[1, 7]
Open	6.3	2.4	[1, 7]
Cons	6.2	2.2	[1, 7]
Agr	6.5	1.9	[1, 7]
Ems	6.5	2	[1, 7]

To investigate whether we could improve predictions by expanding the training examples from one social media source to another one, we employ the three social media datasets that we explained in Section 4. An interesting difference among the

three datasets is the number of labeled users, from thousands of users in Facebook, to hundreds of vloggers in YouTube, and only tens of Twitter users.

One downside in cross-media learning is that we cannot directly use the specific features related to each dataset for training the models, e.g., audio/video features extracted from vlogs or specific users' activities in Facebook. Thus, to make similar training examples, we focus on the common features that we could extract from these datasets. The common features that we use are gender and the linguistic features as we described in Section 5.1 except for the NRC features. Overall, for cross-media learning we extract 161 non zero features for each dataset.

Since the range of the personality scores in our datasets is different, we first map all the scores to values between $[0, 1]$. For this purpose, we consider the actual score ranges of the relevant questionnaire and then map the values by using $f : [min, max] \rightarrow [0, 1] : x \mapsto \frac{x-min}{max-min}$. Since each dataset has used a different questionnaire for calculating the personality scores, we use normalization to obtain training examples with similar personality scores. Table 11 presents the personality score range and the relevant questionnaire type for the datasets that we use in this study.

Another important factor that we consider for cross-media learning regards personality dimensions. In both YouTube and Twitter datasets we have scores for Emotional Stability, however in Facebook, we have the reverse score which is Neuroticism. Thus, we convert the value in the Facebook dataset from Neuroticism (Nue) to Emotional Stability (Ems) by $Ems = 1 - Neu$.

Figure 2 presents the distribution of the converted personality scores in all three datasets. Note that the range of the scores are between [0,1], however the distributions are different which can affect the performance of the cross-media learning experiments. To evaluate the effect of cross-domain learning, we set up the following experiments:

1. **FaceBook+YouTube→FaceBook** ($\{F + Y\} \rightarrow F$): Expand the training examples of the Facebook dataset with YouTube data and apply the learned model on the Facebook testing examples.
2. **FaceBook+Twitter→FaceBook** ($\{F + T\} \rightarrow F$): Expand the training examples of the Facebook dataset with Twitter data and apply the learned model on the Facebook testing examples.
3. **YouTube+FaceBook→YouTube** ($\{Y + F\} \rightarrow Y$): Expand the training examples of the YouTube dataset with Facebook data and apply the learned model on the YouTube testing examples.
4. **YouTube+Twitter→Twitter** ($\{T + Y\} \rightarrow T$): Expand the training examples of the Twitter dataset with YouTube data and apply the learned model on the Twitter testing examples.
5. **Twitter+FaceBook→ Twitter** ($\{T + F\} \rightarrow T$): Expand the training examples of the Twitter dataset with Facebook data and apply the learned model on the Twitter testing examples.
6. **Twitter+YouTube→YouTube** ($\{Y + T\} \rightarrow Y$): Expand the training examples of the YouTube dataset with Twitter data and apply the learned model on the YouTube testing examples.

In all the above experiments, we expand the training examples of one dataset with training examples of another dataset. For this task, we manually create 10

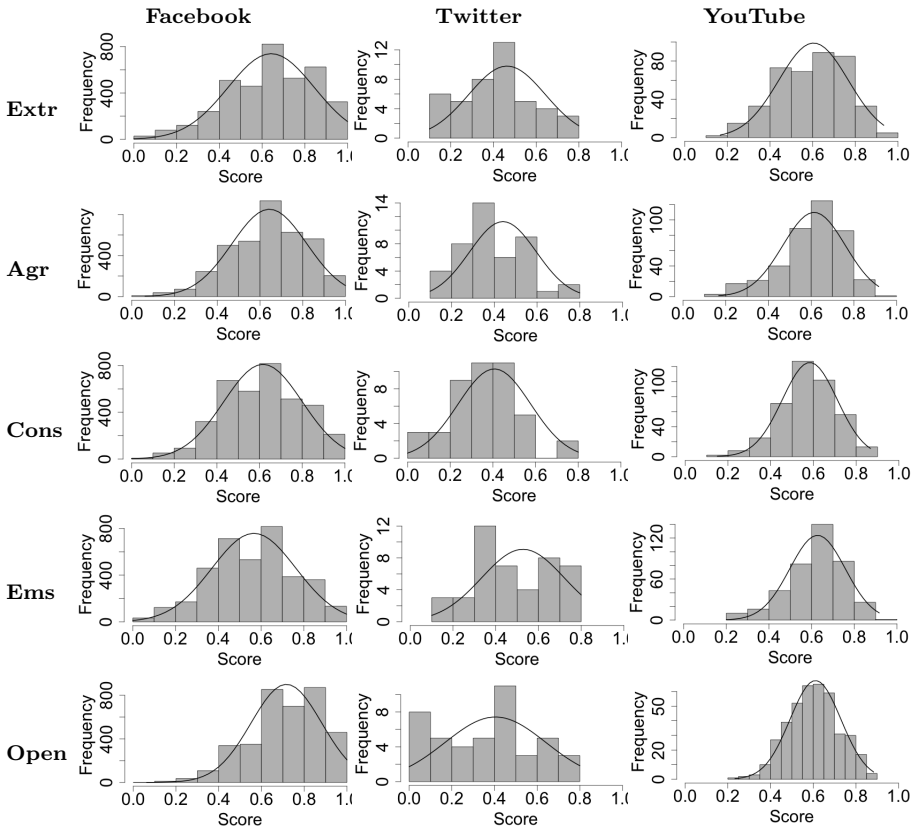

Fig. 2 Distribution of personality scores on five traits, *Extroversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Emotional Stability (Ems) vs. Neuroticism (Neu)*, *Openness (Open)*, in Facebook, YouTube and Twitter datasets. The black curve in each plot presents the normal distribution.

folds out of the first dataset, then each training fold is expanded with the second dataset. The results of this experiment are also averaged over 10 folds.

According to the results presented in the previous section, both $MTSC(DT)$ and $ERCC(DT)$ outperform other learners in all traits across all three datasets compared to other methods. Since the difference between the results of applying $MTSC(DT)$ and $ERCC(DT)$ on all three datasets are not significant, we choose $ERCC(DT)$ as the learning algorithm in this section. Thus, for cross-learning we only focus on the improvement achieved by expanding the training examples using $ERCC$ (DT) as a learner. To compare the results with the situation in which only training examples from the same source are used, we run $ERCC$ (DT) on the three datasets by applying common features and transformed personality scores. Thus, in addition to the above cross-learning experiments, we run the following three experiments:

1. **FaceBook→FaceBook** $(F \rightarrow F)$
2. **YouTube→YouTube** $(Y \rightarrow Y)$
3. **Twitter→Twitter** $(T \rightarrow T)$

Table 12 Root mean square error ($RMSE$) and Coefficient of determination (R^2) results for the personality trait prediction using cross-media learning approaches over Facebook (F), YouTube (Y) and Twitter (T) datasets. The five personality traits are *Extroversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Emotional Stability (Ems) vs. Neuroticism (Neu)*, and *Openness (Open)*. In each column, the lowest error and highest determination are typeset in bold. Significant differences with respect to the baseline ($p < .05$) are marked using $*$. Baseline is the average baseline which is shown by *Avg*.

	Extr		Agr		Cons		Ems		Open	
Approach	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2
Facebook										
Avg	.202		.175		.184		.196		.165	
F→F	**.199**	**2.94**	**.173**	**2.27**	**.179**	**5.36**	**.192**	**4.04**	.163	2.41
{F+Y}→F	**.199**	**2.94**	**.173**	**2.27**	.180	4.30	**.192**	**4.04**	**.162**	**3.60**
{F+T}→F	**.199**	**2.94**	**.173**	**2.27**	.180	4.30	**.192**	**4.04**	.163	2.41
YouTube										
Avg	.163		.147		.129		.130		.120	
Y→Y	.154	10.73	**.121***	**32.25**	.116	19.14	**.117***	**19***	.118	3.31
{Y+F}→Y	**.153**	**11.89**	.130	21.79	.117	17.74	.121	13.37	**.116***	**6.55***
{Y+T}→Y	**.153**	**11.89**	.123	29.99	**.115***	**20.53***	.119	16.21	**.116***	**6.55***
Twitter										
Avg	.179		.159		.175		.198		.236	
T→T	.171*	8.74*	**.151**	**9.81**	.166	10.02	**.183**	**14.58**	**.216**	**16.23**
{T+F}→T	**.170***	**9.80***	.156	3.74	**.165**	**11.10**	.186	11.75	.226	9.96
{T+Y}→T	.175*	4.42*	.161	-9.93	.177	-2.30	.184	13.64	.222	11.51

Note that due to the change in size of the feature space and normalization, results of the experiments listed above are different from those presented in Table 10. The experimental results in Table 12 indicate that extending the training examples of similar datasets, namely Twitter and Facebook, is more effective than an extension with a dataset which has a different context, i.e., YouTube vlogs. It is interesting that in case of the Twitter dataset, where we only have 44 users with personality scores, extending the training examples with both Facebook and YouTube examples indicates no improvement over the training examples of the same source. Besides the Twitter dataset, for the case of YouTube and Facebook, we also gain little or no improvement using the training examples of other sources. These results indicate that the context and respectively the users of these social media sites are different, which is in line with the distribution of the personality scores in Figure 2. Besides the context, the way that the personality scores were calculated are different among these datasets (i.e., observed vs. self-reported), which may also influence the performance of the cross-media learners. Furthermore, having more training examples of the same source makes the performance of the learner more stable, therefore for the case of the Facebook dataset, the performance of the learner by extending the examples with both the YouTube and the Twitter datasets do not differ much. These results differ from the results of cross-media learning of Farnadi et. al in [16], where the performance of the learner was improved by an extension of the training examples. These results suggest that the success of cross-media learning is very dependent on the similarity of two data sources w.r.t. the distribution and calculation of the personality scores.

8 Discussion, Conclusion and Future Directions

In this study, we performed a comparative analysis of state-of-the-art computational personality recognition methods on a varied set of social media ground truth data from Facebook, Twitter and YouTube. We attempted to address three research questions as follows.

(1) Should personality prediction be treated as a multi-label prediction task (i.e., all personality traits of a given user are predicted at once), or should each trait be identified separately? We leveraged a variety of univariate (i.e., decision tree and support vector machine) and multivariate regression techniques (i.e., multi-target stacking, ensemble of regressor chains, and multi-objective random forests) as presented in Section 5.2. When using these learners on the three different datasets, decision tree models mostly outperformed support vector machine models, while multivariate regression learners with decision tree as a base learner often outperformed the univariate regression ones. The differences between univariate and multivariate models were not significant though. Overall the best performing models for this task are the Multi-Target Stacking Corrected ($MTSC$) model and the Ensemble of Regressor Chains Corrected ($ERCC$) model by using a decision tree as a base learner.

(2) Which predictive features work well across different on-line environments? To address this question, we utilized different content-based features (e.g., linguistic features such as LIWC) and context-based features (e.g., audio and video features extracted from vlog videos) in each dataset. We analyzed the correlation between features and personality traits in Section 6.1. We collected the common correlated features with traits among three datasets. From 166 common features for five traits, only 15 common correlations were found. These results suggested that it may not be possible to generalize the correlation between features and the personality traits, as it may vary depending on the underlying data.

Moreover, we measured the performance of the models using different feature sets in addition to the corresponding correlated subset of features. For the YouTube and Facebook datasets, among different feature sets, the LIWC feature set outperformed others for predicting the personality scores of all traits. From the results using both the original feature set and the corresponding correlated feature set, we concluded that selecting features and only using correlated features does not necessarily increase the performance of the learner, however by reducing the size of the feature space we are able to increase the efficiency of the algorithm. Due to the large number of social media users, there is a need to explore efficient models with high performance. Thus, exploring the smallest feature set without loosing the performance in predicting personality traits is an interesting future direction. Furthermore, in this study, we considered Pearson and Spearman correlation as a feature selection approach, however investigating other measures for computing the correlations between features and the personality trait such as *information gain* [34] is an open path to explore.

(3) What is the decay in accuracy when porting models trained in one social media environment to another? To answer this question, we conducted six cross-media learning experiments in which we expand the training examples of one dataset using another dataset. The results were presented in Section 7.

Expanding a model with training examples from another source has not improved the performance of the learner. The context of the data plays a major role

in the success of cross-media learning. Since our YouTube dataset was labeled as perceived personality scores compared to self-reported ones in the case of Facebook and Twitter, a complementary study on the effects of using similar data sources w.r.t. the variation among users and the method for collecting the personality scores in cross-media learning remains a topic for future work.

References

1. Aharony, N., Pan, W., Ip, C., Khayal, I., Pentland, A.: Social fmri: Investigating and shaping social mechanisms in the real world. Pervasive and Mobile Computing **7**(6), 643–659 (2011)
2. Aran, O., Gatica-Perez, D.: Cross-domain personality prediction: from video blogs to small group meetings. In: Proceedings of the 15th ACM on International conference on multimodal interaction, pp. 127–130. ACM (2013)
3. Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., Stillwell, D.: Personality and patterns of Facebook usage. In: Proceedings of the 3rd Annual ACM Web Science Conference (Web-Sci), pp. 24–32. ACM (2012)
4. Back, M.D., Stopfer, J.M., Vazire, S., Gaddis, S., Schmukle, S.C., Egloff, B., Gosling, S.D.: Facebook profiles reflect actual personality, not self-idealization. Psychological Science (2010)
5. Bai, S., Hao, B., Li, A., Yuan, S., Gao, R., Zhu, T.: Predicting Big Five personality traits of microblog users. In: Proceedings of the IEEE/WIC/ACM WI-IAT, vol. 1, pp. 501–508 (2013)
6. Biel, J., Gatica-Perez, D.: The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. Multimedia, IEEE Transactions on **15**(1), 41–55 (2013)
7. Biel, J.I., Aran, O., Gatica-Perez, D.: You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. In: Proceedings of the ICWSM (2011)
8. Blockeel, H., Raedt, L.D., Ramon, J.: Top-down induction of clustering trees. In: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 55–63 (1998)
9. Cantador, I., Fernández-Tobías, I., Bellogín, A., Kosinski, M., Stillwell, D.: Relating personality types with user preferences in multiple entertainment domains. In: Proceedings of the EMPIRE (2013)
10. Celli, F., Lepri, B., Biel, J.I., Gatica-Perez, D., Riccardi, G., Pianesi, F.: The workshop on computational personality recognition 2014. In: Proceedings of the ACM International Conference on Multimedia, pp. 1245–1246. ACM (2014)
11. Celli, F., Rossi, L.: The role of emotional stability in Twitter conversations. In: Proceedings of the Workshop on Semantic Analysis in Social Media. Association for Computational Linguistics, pp. 10–17 (2012)
12. Costa, P.T., McCrae, R.R.: The Revised NEO Personality Inventory (NEO-PI-R). The SAGE Handbook Of Personality Theory And Assessment **2**, 179–198 (2008)
13. Counts, S., Stecher, K.: Self-presentation of personality during online profile creation. In: International AAAI Conference on Weblogs and Social Media (2009)
14. Farnadi, G., Sitaraman, G., Rohani, M., Kosinski, M., Stillwell, D., Moens, M., Davalos, S., De Cock, M.: How are you doing? Emotions and personality in Facebook. In: Proceedings of the EMPIRE, pp. 45–56 (2014)
15. Farnadi, G., Sushmita, S., Sitaraman, G., Ton, N., De Cock, M., Davalos, S.: A Multivariate Regression Approach to Personality Impression Recognition of Vloggers. In: Proceedings of the WCPR, pp. 1–6 (2014)
16. Farnadi, G., Zoghbi, S., Moens, M., De Cock, M.: Recognising personality traits using Facebook status updates. In: Proceedings of the WCPR, pp. 14–18 (2013)
17. Gill, A.J., Oberlander, J., Austin, E.: Rating e-mail personality at zero acquaintance. Personality and Individual Differences **40**(3), 497–507 (2006)
18. Giota, K.G., Kleftaras, G.: The role of personality and depression in problematic use of social networking sites in greece. Journal of Psychosocial Research on Cyberspace **7**(3) (2013)
19. Golbeck, J., Robles, C., Edmondson, M., Turner, K.: Predicting personality from twitter. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, pp. 149–156. IEEE (2011)

20. Golbeck, J., Robles, C., Turner, K.: Predicting personality with social media. In: CHI'11 Extended Abstracts on Human Factors in Computing Systems, pp. 253–262. ACM (2011)
21. Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., Gough, H.G.: The international personality item pool and the future of public-domain personality measures. Journal of Research in Personality **40**(1), 84–96 (2006)
22. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The Journal of Machine Learning Research **3**, 1157–1182 (2003)
23. Hagger-Johnson, G., Egan, V., Stillwell, D.: Are social networking profiles reliable indicators of sensational interests? Journal of Research in Personality **45**(1), 71 – 76 (2011)
24. Hall, M.A.: Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato (1999)
25. Hu, R., Pu, P.: Enhancing collaborative filtering systems with personality information. In: Proceedings of the ACM RecSys, pp. 197–204 (2011)
26. Hughes, D.J., Rowe, M., Batey, M., Lee, A.: A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. Computers in Human Behavior **28**(2), 561–569 (2012)
27. Iacobelli, F., Culotta, A.: Too Neurotic, Not Too Friendly: Structured Personality Classification on Textual Data. In: Proceedings of the Workshop on Computational Personality Recognition, AAAI Press, Melon Park, CA, pp. 19–22 (2013)
28. John, O.P., Srivastava, S.: The Big Five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of Personality: Theory and Research **2**, 102–138 (1999)
29. Jolliffe, I.: Principal component analysis. Wiley Online Library (2002)
30. Kocev, D., Vens, C., Struyf, J., Džeroski, S.: Ensembles of multi-objective decision trees. In: Proceedings of the ECML, pp. 624–631 (2007)
31. Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., Graepel, T.: Manifestations of user personality in website choice and behaviour on online social networks. Machine Learning pp. 1–24 (2013)
32. Kosinski, M., Stillwell, D.J., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy Of Sciences (PNAS) **110**, 5802–5805 (2013)
33. Lambiotte, R., Kosinski, M.: Tracking the Digital Footprints of Personality. Proceedings of the Institute of Electrical and Electronics Engineers (PIEEE) (2014)
34. Lee, C., Lee, G.G.: Information gain and divergence-based feature selection for machine learning-based text categorization. Information Processing & Management **42**(1), 155–165 (2006)
35. Lee, K.M., Nass, C.: Designing social presence of social actors in human computer interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '03, pp. 289–296. ACM (2003)
36. Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research **30**, 457–501 (2007)
37. Mohammad, S., Zhu, X., Martin, J.: Semantic role labeling of emotions in tweets. In: Proceedings of the WASSA, pp. 32–41 (2014)
38. Mohammad, S.M., Kiritchenko, S.: Using nuances of emotion to identify personality. arXiv preprint arXiv:1309.6352 (2013)
39. Nguyen, T., Phung, D.Q., Adams, B., Venkatesh, S.: Towards discovery of influence and personality traits through social link prediction. In: Proceedings of ICWSM, pp. 566–569 (2011)
40. de Oliveira, R., Karatzoglou, A., Cerezo, P.C., de Vicuña, A.A.L., Oliver, N.: Towards a psychographic user model from mobile phone usage. In: Proceedings of the International Conference on Human Factors in Computing Systems, CHI, pp. 2191–2196 (2011)
41. Oliveira, R.D., Cherubini, M., Oliver, N.: Influence of personality on satisfaction with mobile phone services. ACM Transactions on Computer Human Interaction **20**(2), 10:1–10:23 (2013)
42. Ozer, D.J., Benet-Martinez, V.: Personality and The Prediction of Consequential Outcomes. Annual Review of Psychology **57**, 401–421 (2006)
43. Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H., Seligman, M.E.P.: Automatic personality assessment through social media language. Journal of Personality and Social Psychology (JPSP) (2014)
44. Pennebaker, J.W., King, L.A.: Linguistic styles: language use as an individual difference. Journal of personality and social psychology **77**(6), 1296 (1999)

45. Polzehl, T., Moller, S., Metze, F.: Automatically assessing personality from speech. In: Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on, pp. 134–140. IEEE (2010)
46. Quercia, D., Kosinski, M., Stillwell, D., Crowcroft, J.: Our Twitter profiles, our selves: Predicting personality with Twitter. In: Privacy, Security, Risk and Trust (passat), 2011 IEEE Third International Conference on Social Computing (socialcom), pp. 180–185. IEEE (2011)
47. Quercia, D., Lambiotte, R., Kosinski, M., Stillwell, D.J., Crowcroft, J.: The personality of popular Facebook users. In: Proceedings of the Conference on Computer Supported Cooperative Work, pp. 955–964 (2012)
48. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2014). URL `http://www.R-project.org`
49. Rammstedt, B., John, O.P.: Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of research in Personality **41**(1), 203–212 (2007)
50. Saati, B., Salem, M., Brinkman, W.P.: Towards customized user interface skins: investigating user personality and skin colour. Proceedings of the HCI 2005 **2**, 89–93 (2005)
51. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one **8**(9), e73,791 (2013)
52. Stillwell, D.J., Kosinski, M.: myPersonality Project Website. myPersonality Project (2015). URL `http://mypersonality.org`
53. Tausczik, Y.R., Pennebaker, J.W.: The Psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology **29**, 24–54 (2010)
54. Xioufis, E.S., Groves, W., Tsoumakas, G., Vlahavas, I.P.: Multi-label classification methods for multi-target regression. CoRR (2012)
55. Youyou, W., Kosinski, M., Stillwell, D.J.: Computer-based personality judgements are more accurate than those made by humans. Proceedings of The National Academy of Sciences (PNAS) **112**(4), 1036–1040 (2015)