Cognitive Computation (2021) 13:1198–1214
https://doi.org/10.1007/s12559-021-09866-1

The Big-2/ROSe Model of Online Personality

Towards a Lightweight Set of Markers for Characterizing the Behavior of Social Platform Denizens

Gerardo I. Simari[1,2,3] · **Maria Vanina Martinez**[4,5] · **Fabio R. Gallo**[1,2] · **Marcelo A. Falappa**[1,2]

Received: 28 November 2020 / Accepted: 12 April 2021 / Published online: 29 July 2021

Abstract
The Big-5/OCEAN personality traits model, one of the central approaches to psychometrics, has been shown to have many applications over a variety of disciplines. In particular, correlations have been studied leading to effective characterization of people's behavior, and the model has become notorious for its role in the Cambridge Analytica/Facebook scandal surrounding the 2016 US presidential elections. In this paper, we develop Big-2 (or ROSe, for Relationship to Others and to Self), a model via which the personality of users of online platforms can be studied using a lightweight set of markers focused on online behavior, avoiding the major data privacy pitfalls afflicting approaches based on more powerful models that characterize personal aspects of the human psyche. Evaluation of Big-2's effectiveness is done in two parts: a quantitative evaluation on a specific prediction task and a qualitative one based on an analysis of the different ways in which the Big-2 traits can be derived from online behavior, proposing a general template to guide such efforts. Quantitative results show that our lightweight model can match or surpass the performance of Big-5 in a prediction task, while qualitative results show that it is feasible to implement the model based on the observation of basic online user behavior. Our main result is a general-purpose model that can be used to characterize the personality traits of users of online platforms in an ethical manner. Our proposed model provides a valuable tool to carry out effective and explainable analyses of online personality, avoiding the collection of unnecessary user data that would open the possibility for ethical violations.

Keywords Personality models · Online personality · Cognitive models · Behavior prediction · Explainability and interpretability · Machine learning · Ethical AI

Motivation and Related Work

The analysis of social networks, and the information that flows through them, has been an active topic of research in several disciplines connected to the study of the different aspects of communication among people or entities—sociology, psychology, philosophy, economics, and computer science are some examples. The development and use of *psychometrics*—a field concerned with the measurement of different aspects of the human psyche [1]—has in recent years been in the public stage due to its application in influencing political campaigns around the world, most notably the presidential elections in the USA and the "Brexit vote" in the UK (both occurring in 2016). In those cases, it came to light that social media users were manipulated [2, 3] based on what has now become known as *fake news*; however, the effectiveness of such malicious attempts at disinformation is optimized when a deep personality profile is associated with each individual. In these well-known cases, such profiles were built based on the Big-5/OCEAN model [4], which we will discuss below in detail.

Figure 1 presents the general process (that we assume as context in this paper) of developing applications that involve the use of personality models towards solving a specific real-world problem. The first step involves deriving one or more personality models for a user (such as Big-5, Holland Hexagon), while the second step leverages such models in developing an application to be deployed in some domain. For instance, in our previous work [5] we developed a system to predict basic actions that users take in response to the

✉ Gerardo I. Simari
gis@cs.uns.edu.ar

Extended author information available on the last page of the article

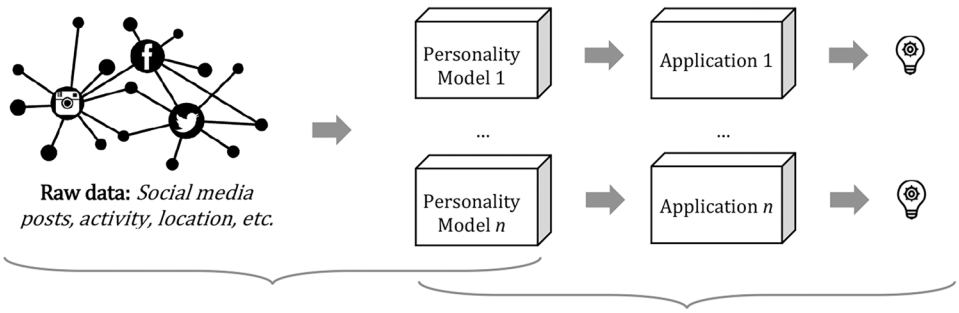

Fig. 1 Overview of the process of developing applications that leverage personality models to deliver insights about users participating in social media platforms—this paper proposes and evaluates one such model, developing prototype implementations for steps 1 and 2 but otherwise considering them as black boxes

content of their Twitter feeds, which involved first deriving a Big-5 personality profile based on their past tweets and then incorporating this profile into a broader set of features to train a machine learning classifier that ultimately delivers the desired prediction. Below we give further examples of other related work that takes this same approach.

We now first discuss research lines that are related to this work, and then summarize the main contributions and organization of the paper.

Related Work

Since the main topic of our work is the characterization of online users towards understanding and predicting their behavior, we will discuss research from different fields—such as politics, economics, and health—focusing on aspects that share this goal. We first discuss efforts for making predictions and detections based on general social media content, and then focus especially on such tasks leveraging personality models.

Predictions and Detections based on General Content There is a large body of research and development that seeks to make predictions based on signals available in social media platforms. In the health arena, there is work showing how certain psychological disorders can be detected based on data from social media profiles, such as addiction-like symptoms related to social platforms [6] or obsessive-compulsive and bipolar disorders [7, 8]. In economics, social media activity is often used as a means to gauge the mood of a group of people (such as a society), or to predict events of interest. Examples of rather early attempts are the work of [9], which shows how collective mood derived from Twitter activity correlates with the stock market, or correlations between tweet rates about certain topics and box office revenue [10]. An interesting aspect of the latter is that the authors show that performance improves when tweets are also analyzed for sentiment.

Politics is another area in which social media activity is central, as already mentioned above. A clear application is predicting the outcomes of elections, as done in [11, 12] using Twitter mentions of political parties, social circles, and additional demographic information. Other efforts include [13], where Twitter discussions surrounding elections are analyzed using tools based on bot detection, network science, and sentiment analysis, or [14], where the authors show how Twitter analytics can be used to predict not only outcomes of assembly elections, but also vote and seat shares.

In a related effort, the recent work of [15] shows how analysis of Twitter feeds can be combined with machine learning tools to detect cyberbullying; similarly, [16] shows how personality type can also be leveraged towards this end. The work of [17] includes a survey of work on predictions using social media content.

Leveraging Personality Models The theory and practice of personality modeling is too extensive to adequately survey in this context; we provide a brief recap here but refer the interested reader to [18] for a deeper treatment of the historical roots and modern evolution of this area of study; references to the original proposals of the main models can also be found therein.

The oldest tools that have been developed to model personality and emotion date at least to the ancient Greeks with the proposal of the Four Temperaments (or Humors) in connection with the four natural elements (these, in turn, are likely to be based on even older traditions). The development of Psychology in the early twentieth century continued this line of thinking, and Jung proposed his eight personality types as arising from the combination of two General Attitude types (introverted and extraverted) and four Functional Types (thinking, feeling, sensation, and intuition). A refinement of this model, including a principal and auxiliary function, naturally leads to sixteen types. The Jungian model later served as the basis for the Myers-Briggs Type Indicator proposed in 1942, one of the better-known models still in use almost eight decades later. This model also divides personality into 16 possible types, arising from four dimensions according to which extreme is preferred: extraversion

vs. introversion, sensing vs. intuition, thinking vs. feeling, and judging vs. perceiving. Holland codes were later developed in the late 1950s to model vocational personalities and working environments, which divide types into doers, thinkers, creators, helpers, persuaders, and organizers. This division gives rise to the RIASEC Hexagon (also known as the Holland hexagon), placing the different types on a Cartesian plane divided by two main dimensions: things vs. people, and data vs. ideas. Other models include the HEXACO personality structure, the DISC assessment, and the Big-5/OCEAN model, which we discuss in detail in the next section.

We now continue with a discussion of how these models are typically used, and then focus on the main computational approaches developed to obtain a personality model for a specific user in an automated fashion. The body of work related to computational personality models typically falls into one of two categories: those attempting to predict personality type from content, and those leveraging already available personality types as features for making better predictions—these correspond respectively to Steps 1 and 2 in Fig. 1. For instance, [19] seeks to predict users' Big-5 scores based on information that is publicly available on their profiles, while [20] applies regression algorithms to predict such values from user behaviors. A recent survey of related techniques can be found in [21]; below we discuss recent trends in the application of biologically-inspired computing and AI approaches to personality detection.

The other category is of more interest for our purposes; in the context of the Big-5/OCEAN model (which assigns a value to a person's openness, conscientiousness, extraversion, agreeableness, and neuroticism) much work has been dedicated to identifying which traits are relevant when attempting to predict specific behaviors. A clear example of this is the work of [22], which showed that high values of *openness* and low values of *neuroticism* correlate with a more favorable response to targeted advertisements. Moreover, *extraversion* and *openness* are positively related to social media use, while emotional stability was a negative predictor [23]; see [24] for other works obtaining results along these lines. Our previous work [5], discussed in greater detail below, also belongs to this group.

Closely related to the works discussed above, adding sentiment and personality analysis has been shown to improve the accuracy of predictions of consumer behaviors [25, 26], depression [27], Internet-related pathologies [28], political party affiliations [29], and socialization [30, 31].

Bio-inspired Computing Approaches to Personality Detection and Predictions In recent years, the widespread availability of digital footprints—such as those arising from activity on social platforms—has led to research and development of a variety of approaches to detection and prediction of personality traits based on machine learning approaches. State-of-the-art tools [32] focus on language-based trait prediction combining both psycholinguistic features and language model embeddings. Other approaches also consider non-linguistic features such as those arising from smartphone use and activity on social media [33, 34]. Finally, recent approaches have focused on multi-task learning frameworks for predicting both social media users' personality traits and emotion [35], attention-based deep models for sentiment analysis [36, 37], and stacked ensemble models for related problems like predicting bipolar disorders [7] and sentiment intensity [38].

In this paper, we will generally consider the step deriving personality traits from raw data (Step 1 in Fig. 1) as a black box, focusing instead on the development of a novel set of traits and analyzing its properties. In the next section we discuss the particulars of this contribution, and how we evaluate it both quantitatively and qualitatively. For a recent survey on this topic, please see [39].

Contributions and Organization of this Paper

The main contribution of this paper is the development and evaluation of Big-2/ROSe, a new model for characterizing users of online platforms that focuses on two main objectives centered on four criteria, which we discuss in two pairs:

(i) *Privacy and Proportionality*: while models like Big-5 are powerful in that they characterize many facets of an individual's personality (30, as shown in Fig. 2), they also facilitate overstepping as done by Cambridge Analytica as mentioned above. Essentially, this is because the Big-5 model includes facets that characterize very personal aspects of a user's psyche, like liberalism, morality, and vulnerability, which paves the way towards finding the best strategies for manipulation. Our model is designed as a *lightweight* alternative to Big-5, focusing on less invasive facets that are germane only to their online behavior.

(ii) *Interpretability and Explainability*: the direct application of general-purpose models (like Big-5) has several drawbacks, such as the fact that their complexity is often addressed by applying tools based on machine learning that are not designed to offer access to their inner workings, nor an explanation along with their answers.

To address this, we propose implementation schemes for each facet of our model, allowing to trace back the reasons behind each assigned value. Though this is in principle possible for more complex models, lightweight ones have a clear advantage given their reduced set of traits and facets. Figure 1 illustrates this aspect: if the raw data used in Step 1

Openness (O)		
Consistent	**Adventurousness**	Adventurous
Unconcerned with art	**Artistic interests**	Appreciative of art
Dispassionate	**Emotionality**	Emotionally aware
Down-to-earth	**Imagination**	Imaginative
Concrete	**Intellect**	Philosophical
Respctful of authority	**Liberalism**	Authority-challenging
Conscientiousness (C)		
Content	**Achievement-striving**	Driven
Bold	**Cautiousness**	Deliberate
Carefree	**Dutifulness**	Dutiful
Unstructured	**Orderliness**	Organized
Intermittent	**Self-discipline**	Persistent
Self-doubting	**Self-efficacy**	Self-assured
Extraversion (E)		
Laid-back	**Activity level**	Energetic
Demure	**Assertiveness**	Assertive
Solemn	**Cheerfulness**	Cheerful
Calm-seeking	**Excitement-seeking**	Excitement-seeking
Reserved	**Friendliness**	Outgoing
Independent	**Gregariousness**	Sociable
Agreeableness (A)		
Self-focused	**Altruism**	Altruistic
Contrary	**Cooperation**	Accommodating
Proud	**Modesty**	Modest
Coompromising	**Morality**	Uncompromising
Hard-hearted	**Sympathy**	Empathetic
Cautious of others	**Trust**	Trusting of others
Neuroticism/Emotional Range (N)		
Mild-tempered	**Anger**	Fiery
Self-assured	**Anxiety**	Prone to worry
Content	**Depression**	Melancholy
Self-controlled	**Immoderation**	Hedonistic
Confident	**Self-consciousness**	Self-conscious
Calm under pressure	**Vulnerability**	Susceptible to stress

Fig. 2 Overview of the Big-5/OCEAN personality traits as implemented in the IBM Watson Personality Insights service; *center*: facet name, *left*: description for low value, *right*: description for high value

is only available to the tools used to derive the personality model, then the richness of the features afforded by such a model will be the limiting factor of the extent to which the application using them can go. We return to this discussion in Section 4, where we examine a concrete use case as part of the qualitative evaluation of our model.

The remaining contributions are centered on the evaluation of the proposed model. First, we carry out a quantitative evaluation to show the effectiveness of one possible implementation on a prediction task. Then, we perform a qualitative evaluation of the model's features in terms of data privacy, proportionality, interpretability, and explainability.

This work is part of a broader line of research that seeks to build a "map" of user types in social platforms describing how they react to different kinds of content, with the goal of understanding the principles underlying the flow of information in these media. We believe that this is a key element in the fight against malicious content—which is also sometimes called *pathogenic social media*—that lies at the root of trolling and bullying campaigns, misinformation, and other kinds of manipulation. This line of work began in [40] and later focused on belief dynamics [41, 42] with the development of the *network knowledge base* model. Most recently, in [5] we adapted a simple version of that model for solving the task of predicting the reaction of users given the content of their Twitter feeds and Big-5 personality traits. In this paper, we develop Big-2/ROSe, a novel personality model especially geared towards describing *online* personality; the quantitative evaluation mentioned above is done in the same setting used in [5]. Finally, note that the idea of abstracting personality traits is not a novel one, and several approaches have adopted it in the pursuit of understanding personality, arriving at models that are also sometimes referred to as "Big-2". An early example is [43], and [44] later also studied meta traits from different points of view. In both cases—as well as others in the literature—the authors identify a hierarchical relationship very much in line with our proposal, albeit taking a non-computational stance.

The rest of this paper is organized as follows: Section 2 first provides an overview of the Big-5/OCEAN model and then presents the set of traits and facets that comprise Big-2/ROSe, including a baseline algorithm for implementing it based on the former. Section 3 is dedicated to a quantitative evaluation showing that machine learning classifiers including personality types from the baseline implementation of Big-2 already perform comparably well in comparison to those using the full Big-5 model in a basic user behavior prediction task, and in some cases significantly better. Section 4 then focuses on a qualitative evaluation of the benefits of the Big-2 model in comparison with more complete models like Big-5, Needs, Values, or HEXACO, and then goes on to discuss several approaches to implementing the model in a way in which such benefits can be reaped. Finally, Section 5 includes closing remarks and discusses future work.

The Big-2/ROSe Model

We now present our Big-2 model of *online* personality, which we also refer to as ROSe (for Relationship to Others and to Self). The model is designed to be a lightweight version of the well-known Big-5 model, also known as OCEAN (for Opennes, Conscientiousness, Extroversion, Agreeableness, and Neuroticism). It was developed by carrying out an exhaustive analysis of the Big-5/OCEAN model's traits and facets, from a functional point view. More specifically, we evaluated each facet in terms of its projection onto aspects that manifest themselves in social platforms, keeping in mind the central goals of privacy, proportionality, interpretability, and explainability described above.

We first provide an overview of Big-5, and then move on to the presentation of our model.

Overview of the Big-5/OCEAN Model

In this paper we consider the variant of Big-5 as implemented in IBM Watson's Personality Insights service [24]. The model is structured into five central *traits*, each of which is divided into six *facets*—Fig. 2 presents a summary of the model[1], specifying basic interpretations for low and high values of each facet. The alternative name "OCEAN" arises as an acronym comprised of the first letter of each trait.

Note that "high" and "low" values are not associated with "good" and "bad" properties—they are simply values in a spectrum, such as in the case of *Artistic interests*, which ranges from "unconcerned with art" (low values) to "appreciative of art" (high values)[2]. A typical way of deriving a Big-5 *personality type* is to divide each trait's value into either *high* or *low*; interpreting the result as a binary vector, a value in the interval [1,32] is then obtained.

Two Traits Grouping Six Facets

We now describe the Big-2/ROSe model; as mentioned in Section 1, it is designed to be a lightweight version of Big-5/OCEAN. The two traits in total comprise six facets—recall that the goal of the model is to capture *online* behavior, and therefore aims to be a less powerful characterization compared to Big-5.

- **Relationship to Self (RS)**: This trait comprises how a person behaves with respect to her/himself in online platforms. High values tend to characterize cheerful, energetic, open people who are typically comfortable with who they are, while low values correspond to the other end of the spectrum: more private people with lower levels of energy who are less comfortable with themselves and are possibly suffering from some kind of depression. RS is divided into the following more specific facets or sub-traits:
  - *Self Worth* (F1): Feature related to happiness, anxiety, desires, and self-worth in general.
  - *Structure* (F2): Characterizes organization, consistency, attitude towards satisfying own needs, and openness to explore new things.
  - *Activity Level* (F3): Captures a person's focus on privacy, energy, assertiveness, and risk.
- **Relationship to Others (RO)**: This trait captures the dual of RS: how a person relates to others online. High values indicate cooperative, fiery, empathetic people who tend to behave well with others, while low values characterize more self-centered individuals who are less diplomatic. This dimension is subdivided into the following facets:
  - *Selflessness* (F4): Describes a person's tendencies toward focus on others, cooperation, and trust.
  - *Smoothness* (F5): Describes tendencies toward friendliness, diplomacy, and empathy.
  - *Emotionality* (F6): Captures temper, passion, depression, pleasure, discipline, optimism, and desire to share joy and positiveness.

[1] This table partially reproduces the content of the table available at https://watson-developer-cloud.github.io/doc-tutorial-downloads/personality-insights/Personality-Insights-Facet-Characteristics.pdf

[2] For a more detailed discussion on the interpretation of numeric values, see https://cloud.ibm.com/docs/services/personality-insights?topic=personality-insights-numeric

	RO *High*	RO *Low*
RS *High*	Confident, Open, Energetic, Risk-taker (self) Empathic, Passionate, Cooperative, Diplomatic	Confident, Open, Energetic, Risk-taker (self) Self-involved, Ill-managed, Apathetic, Mild-tempered
RS *Low*	Self-conscious, Introvert, Laid-back, Cooperative, Empathic, Passionate, Diplomatic	Self-conscious, Introvert, Laid-back, Self-involved, Apathetic, Ill-managed, Mild-tempered

Fig. 3 Summary of main characteristics associated with combinations of *high* and *low* values for the two ROSe facets

High levels of the RS trait involve high levels of self-worth, structure, and activity level, while low levels of the three facets define low levels of the trait. On the other hand, for the RO trait, high levels involve high levels of smoothness and emotional expressiveness, and selflessness. Conversely, low levels are defined by low levels of selflessness, smoothness, and emotional expressiveness. Figure 3 provides a general description of individuals that fall within the more extreme values of each trait. As we will discuss below, contemplating a range of values wider than just high vs. low allows to characterize a larger number of personality types, just like in OCEAN.

A Proxy Instantiation of Big-2/ROSe based on Big-5 Facets

Our proposed model is very flexible and can be implemented in many ways. In the next section, we report on the results of a quantitative evaluation of a baseline implementation that directly relies on combining values from Big-5 personality profiles; the goal of that evaluation is to show that the model performs well on a task seeking to predict how users will react to the content of their Twitter feeds. Furthermore, we refer to this as a *proxy* or a *baseline* instantiation because it is meant to show one use case for the model; as we discuss in Section 4, which describes the results of a qualitative evaluation, more sophisticated implementations are likely to afford even better results and also provide better tools for interpretation of this kind of predictions.

For each Big-2 trait, we associate a set of Big-5 facets whose values can be used to approximate the value for the new model; this assignment was done as follows.

- **RS Trait:** Cheerfulness, Self-efficacy, Anxiety, Immoderation, Self-consciousness, Orderliness, Altruism, Adventurousness, Friendliness, Activity level, Assertiveness, Excitement-seeking.
- **RO Trait:** Cooperation, Sympathy, Cautiousness, Anger, Emotionality, Self-discipline, Depression, Self-efficacy, Cheerfulness, Self-consciousness, Immoderation.

We tested several variants of this procedure, but for reasons of space we present the one yielding the best performance in the experiments (cf. Section 3.2).

In practice, in order to compute an actual value for each Big-2 trait, we consider the value of each Big-5 facet, according to the percentile returned by the IBM API, for each user. Assume an interval $(\ell, u) \subseteq [0, 1]$, we consider as facets with *tail* values, those Big-5 facets with percentile values outside the interval (i.e., $\leq \ell$ or $\geq u$); intuitively, they represent exceptionally high or low scores for a given sub-trait in an individual's personality profile. For each Big-2 trait, we first filter from the Big-5 facets associated to it, those that have a tail value. If the resulting set of traits is non-empty, the value for the Big-2 trait is the mean of all remaining Big-5 values; otherwise, we revert to the full set of Big-5 features comprising the Big-2 trait and take the mean of all such values. Figures 4 illustrates the mapping from Big-5 facets to Big-2 traits, indicating also to what Big-5 trait each of them belongs. Figure 5 illustrates an example of how the mapping works in practice.

Note that in this baseline approach we map Big-5 facets to Big-2 traits—that is, we do not take into account facets F1–F6 described above. Alternatively, for each

	Big-5 Facets	O	C	E	A	N
Relationship to Self (RS)	Cheerfulness			×		
	Self-efficacy		×			
	Anxiety					×
	Immoderation					×
	Self-consciousness					×
	Orderliness		×			
	Altruism				×	
	Adventurousness	×				
	Friendliness			×		
	Activity level			×		
	Assertiveness			×		
	Excitement-seeking			×		
Relationship to Others (RO)	Cooperation				×	
	Anger					×
	Cautiousness		×			
	Depression					×
	Immoderation					×
	Sympathy				×	
	Self-discipline		×			
	Self-efficacy		×			
	Self-consciousness					×
	Emotionality	×				

Fig. 4 Summary of our proxy instantiation of Big-2/ROSe by deriving its traits using combinations of Big-5/OCEAN facets. Crosses indicate to what Big-5 trait each facet belongs

	Big-5 Facets	%tile	Tail?	Value
RS	Cheerfulness	0.209		0.78
	Self-efficacy	0.951	×	
	Anxiety	0.464		
	Immoderation	0.000	×	
	Self-consciousness	0.294		
	Orderliness	0.280		
	Altruism	0.968	×	
	Adventurousness	0.977	×	
	Friendliness	0.600		
	Activity level	0.931		
	Assertiveness	0.987	×	
	Excitement-seeking	0.523		
RO	Cooperation	0.235		0.65
	Anger	0.274		
	Cautiousness	0.685		
	Depression	0.208		
	Immoderation	0.000	×	
	Sympathy	0.998	×	
	Self-discipline	0.756		
	Self-efficacy	0.951	×	
	Self-consciousness	0.294		
	Emotionality	0.197		

Fig. 5 Example of derivation of values for Big-2 based on Big-5 in our baseline implementation

of these facets we can try and identify a set of features/variables that can be measured quantitatively from the subjects' data directly. For instance, facet F4 (selflessness) could be measured in terms of the extent and frequency to which the subject interacts with his/her connections, the frequency with which subjects enter into debates with them, the vocabulary and tone (sentiment) they use when/if they do, etc. On the other hand, F5 (smoothness) can be assessed by analyzing the vocabulary and sentiment used when responding to people or reacting to posts, and whether they respond to offensive or impolite (praiseful, respectively) posts directed to them or that mention them or their connections. In Section 4, we evaluate the benefits of such an approach in terms of privacy, proportionality, interpretability, and explainability of the model, including a concrete example to ground the analysis.

Comparing the Effectiveness of Big-2 vs. Big-5 in one Prediction Tool

In this section, we describe an empirical evaluation designed to test the usefulness of the Big-2 model in a basic prediction task related to social media: *given past behavior and the current context, what can we expect a given user to do in the following time step?* We first provide details of the setup and then discuss the results. The experiment was conducted based on those described in our previous work [5], in which the predictive power of Big-5 personality type was shown to be a significant feature in a machine learning classifier; the contents of the following subsection are therefore based on our description in [5].

Experimental Setup

We now discuss the main aspects of the setup[3]. The experiments were run on a computer with an AMD A8-7650K Radeon R7 processor at 3.3GHz and 4GB of RAM, using Python 3.6.4 (sklearn v0.2 library).

Features.

We selected the following set of basic features; the numbers correspond to the enumeration in Fig. 6 (top):

- (1) *Personality Type.* We use values computed according to the Big-5/OCEAN model as well as the Big-2/ROSe model, the latter calculated as described in Section 2 (baseline instantiation, with tail values 0.05 and 0.95).
- (2) *Time Step/Interval.* We discretize time into intervals; for this study, we refer to a collection of intervals selected from the most recent up to a certain point in the past as the *context* considered for the prediction task. We use k to denote the number of such intervals that make up the context.
- (3) *Predominant Sentiment.* We analyze the overall sentiment (or tone) present in each tweet and classify it into *positive*, *negative*, or *neutral*. Then, for a given context, we consider the overall tone that the user was exposed to, and refer to this as the context's *predominant sentiment*.
- (4)(5) *Sentiment Distribution.* As a refinement of the predominant sentiment feature, we also consider the distribution of positive and negative sentiment as a measure of the strength of the predominant sentiment. This is discretized into four intervals: [0, 25), [25, 50), [50, 75), and [75, 100], indicating the percentage of items in the feed with positive sentiment, and an additional such value for negative sentiment.

Given these features, the prediction task seeks to decide whether a user will either *take action* or not, where "action" refers to the generation of content such as using a *new* hashtag (one that does not appear in the current context), reusing a hashtag with the *same* sentiment as the predominant one associated with it, or reusing a hashtag with a *change* in sentiment. Figure 6 illustrates the prediction task and basic classifier setup, as described next.

[3] The code used for these experiments is available at: https://github.com/fabiorgallo/Big2-OCEAN-Experiment

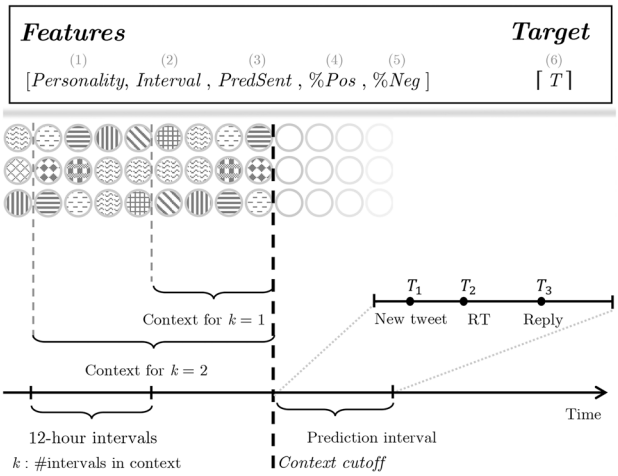

Fig. 6 Experimental setup—*Top*: Overview of the features considered in our model; *Bottom*: Sketch of the prediction task used in the experiments

Dataset. The dataset is comprised of 18,292,721 Twitter posts published between July 15, 2013 and March 25, 2015; we only considered posts written in English (16,780,489). To track user reactions, we focus on tweets containing hashtags, which can be found in 5,107,986 tweets (136,809 distinct hashtags in total). The dataset also contains information on the underlying network of follow/friend relations, which is fundamental in the reconstruction of the feeds for each user. The specific content of the posts is not of relevance for this study; this dataset was originally collected to analyze elections in India[4].

Data Preparation and External Services For each tweet containing a hashtag, we create a post; then, to build the setup illustrated in Fig. 6 (bottom), we prepared each user's *feed* comprised of the tweets posted by users he/she follows. Such feeds are the bases for preparing the *context* used in the prediction task.

In order to obtain a value for the Big-5 *personality type* feature, we used the *Personality Insights* service provided by IBM Cloud[5][6]. This API analyzes text and derives a personality profile of its author, and provides a value for each of the 30 facets in Fig. 2 (both as an absolute value and a percentile, which is calculated based on a population of approximately one million users [24]). The process works as follows: given a user, build one text file by concatenating all his/her posts, and submit it to the API; then, as discussed above, we discretize the values for each dimension into *high* and *low* (or "+" and "–") to obtain a value between 1 and 32. The derivation of values for the Big-2 model is similar, but instead of using *high* and *low* to discretize the values (which would be too coarse for Big-2) we use four levels, thus yielding a value between 1 and 16.

To detect the general tone of a post, we used the PHPInsight tool[7], which yields a value of either *positive*, *negative*, or *neutral*.

Classifier, Hyperparameters, and Other Experimental Parameters As mentioned above, this experiment follows the setup of prior work [5]; we therefore used the same configuration found to yield the best results there: Multinomial Naive Bayes classifier, with hyperparameters *alpha* (smoothing parameter) set to 0.1 and *norm* (second normalization of weights performed) set to *true*. As usual, we use 90% of the available data for training, and the remaining 10% for testing.

The following parameters are also part of our experimental setup:

Number of intervals in the context (k), with possible values 1 or 2.

Spread: There is typically great variability in user behavior; on the one hand there are many users who post very little, and on the other many are very active posters. There is therefore a severe class imbalance when data for all users is considered for the prediction task; in our dataset, the distribution of intervals in which users take action vs. no action is on average approximately 8%/92%. To investigate the effect that such class imbalance has on the performance of our classifiers, we selected users according to a parameter called *spread*; a user will be chosen according to a value of x for this parameter if the difference (in percentage points) between the percentage of intervals for which action was taken vs. no action was taken is *at most* x. So, for instance, for a value of 50, a user with 62% intervals with action and 38% with no action is selected ($62 - 38 \leq 50$). Since intervals are discarded for which nothing is received in a user's feed and no action is taken, the value of k also influences the number of users chosen.

Fig. 7 shows the number of users selected for a variation of the parameter in [50,100] in 5-point increments, for both values of k—the number of users participating in the prediction task (both in training and testing) will be denoted with #Users. As we will see, this quantity has a significant impact on the performance of the classifiers.

Evaluation Metrics To evaluate performance, we adopt the typical metrics of *precision* (ratio of true positives to overall positives, indicating the proportion of the selected elements

[4] We thank V.S.Subrahmanian for sharing the dataset.

[5] https://www.ibm.com/watson/services/personality-insights/

[6] We thank Constanza Caorsi from IBM Argentina for facilitating an academic license for this service.

[7] https://github.com/JWHennessey/phpInsight

Spread (% pts.)	k=1	k=2
50	18	3
55	21	7
60	30	12
65	38	23
70	81	41
75	170	100
80	372	265
85	816	628
90	1,568	1,348
95	2,294	2,213
100	3,048	3,048

Fig. 7 Number of users satisfying the *spread criterion* for different values of percentage points and k

that are relevant), *recall* (ratio of true positives to the sum of true positives and false negatives, indicating the proportion of relevant elements that are selected), F1, and F_β. The latter two are summary metrics that combine precision and recall into a single value. F1 is the harmonic mean of the two values and weights them equally; F_β, on the other hand, assigns weights according to the value of β—the higher the value, the more importance it places on recall (and vice versa). We consider values of $\beta \in \{0.5, 2, 3, 4\}$.

Results

The results are displayed in Figs. 8, 9, and 10. The first two include eight plots each: on the left we show the performance of the OCEAN-based classifier (the first component of the feature vector in Fig 6 (top) is the value in [1,32] corresponding to the OCEAN model), while on the right we have the performance of the ROSe-based one (the first component is a value in [1,16] computed as described above). Each of the top six plots also includes the result of an *ablation* study—the solid line represents the performance of the full classifier, while the dashed line is the performance of the classifier resulting from not considering the personality type component. Figure 10 shows area under the curve (AUC) calculations for the previous two figures; additionally, AUC values for *restricted* scenarios (#Users $\leq$ 100) are included to show how variability of behavior affects the performance of the classifiers.

For $k = 1$, we see an interesting tradeoff between precision and recall for the two approaches; Big-2 gains a substantial boost in precision (18.24% over all users, 31.69% for the focused setting) though recall drops from nearly perfect scores (39.48% and 20.05%, respectively); however, recall remains above 0.6 (0.8, respectively). On the other hand, for $k = 2$ we observe very similar performance for both classifiers, with only minor variations between −0.79% and 0.89% in AUC.

These results show that the Big-2/ROSe model has potential for replacing the Big-5/OCEAN in this kind of prediction tasks. The fact that a significant overall increment in precision and F1 score for $k = 1$ was observed is also promising, since our prior work showed that there is quite some room for improvement in precision after evaluating a wide range of classifiers based on Big-5/OCEAN. Another interesting observation is that this increment is only seen for small context sizes (recall that $k = 1$ means that only the feed for the last 12 hours is considered); this confirms the informal observation that online platform users tend to focus their attention on the most recent content of their feeds.

Limitations of this Study

There are several limitations to the quantitative evaluation we carried out. First of all, the implementation of the Big-2/ROSe model used here is meant to be one approximation of the model, serving as the basis for an initial evaluation. Though the results are encouraging, we expect more involved implementations to have both better performance and greater benefits—this is the topic of the next section, in which we carry out a qualitative evaluation. Furthermore, the same kind of analysis should be performed on other problems and datasets, and an expanded range of values for some of the parameters to further reduce the possibility of confounding variables affecting our results—this is part of ongoing work. Finally, exploring richer tools for sentiment analysis is likely to afford improved performance of tools like the classifiers evaluated here. Sentiment analysis is a fertile research topic; the community is working on many fronts, such as developing tools that are capable of handling complex domains [45, 46] and advanced features such as detection of figurative language [47].

A Qualitative Evaluation of the Big-2 Model

As a complement to the quantitative evaluation described in the previous section, we now report on the results of a *qualitative* evaluation of our proposed model. We will do so by first analyzing it in terms of data privacy and proportionality (Section 4.1) and then how well it supports the highly sought-after qualities of interpretability and explainability (Section 4.2). Figure 11 illustrates a simple example of a social media post made by user U_1 and the exchange it sets off with another user, U_2. This example will be used to illustrate the main points we present in this section.

Issues of Data Privacy and Proportionality

Models such as Big-5, Holland Hexagon, or Need-Values (cf. Section 1.1) aim to understand a wide spectrum of human personality traits, and this process was traditionally done using statistical analyses of the results of a qualitative survey completed by the subjects. Attempts to use this method to

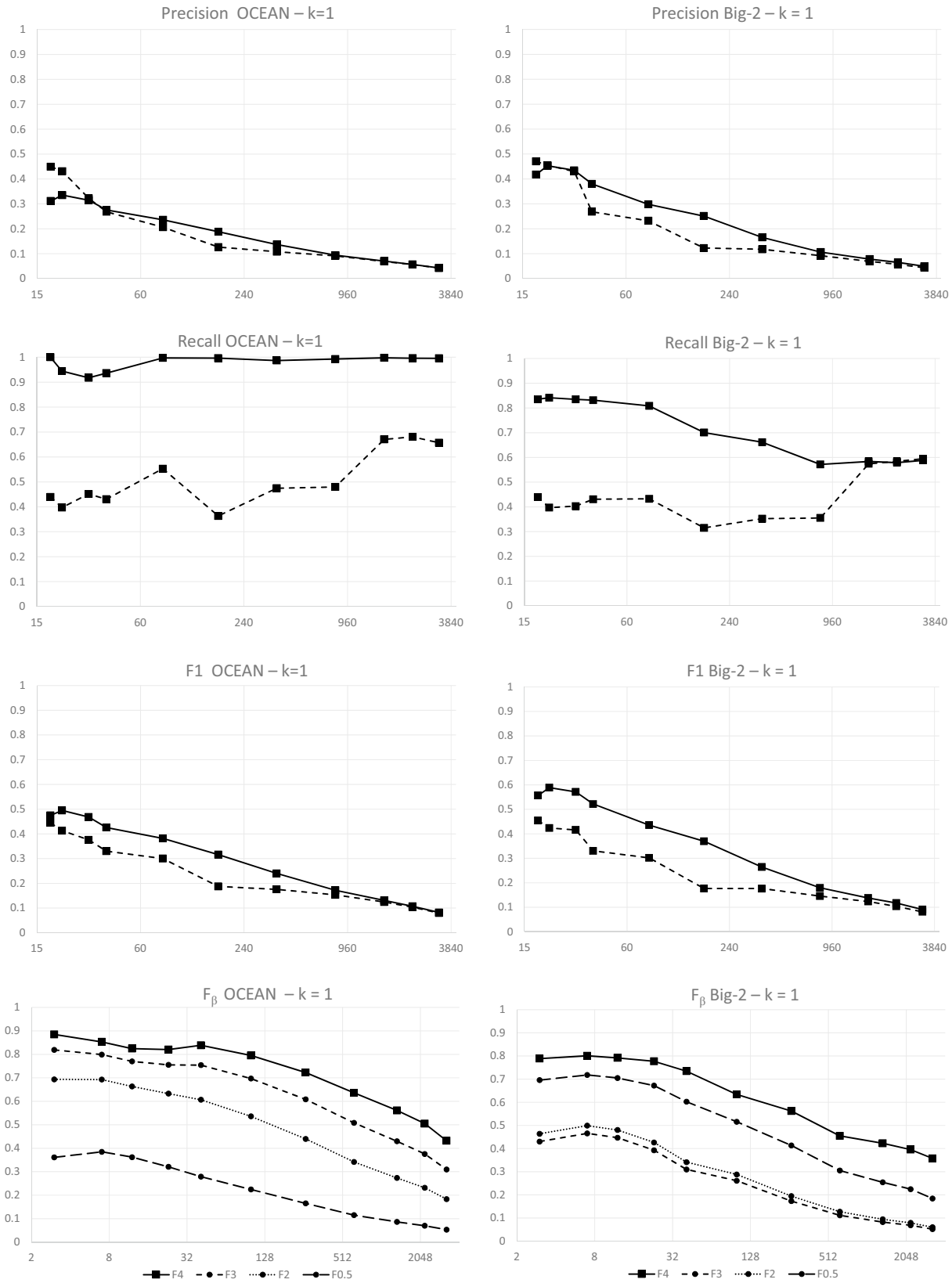

Fig. 8 Performance of classifiers trained using Big-5/OCEAN (left) and Big-2/ROSe (right) for $k = 1$, varying number of users selected for the training phase. Each chart plots the full classifier (solid line) against the same classifier but without including the personality model feature (ablation study). Cf. Figure 10 for pairwise area under the curve comparisons

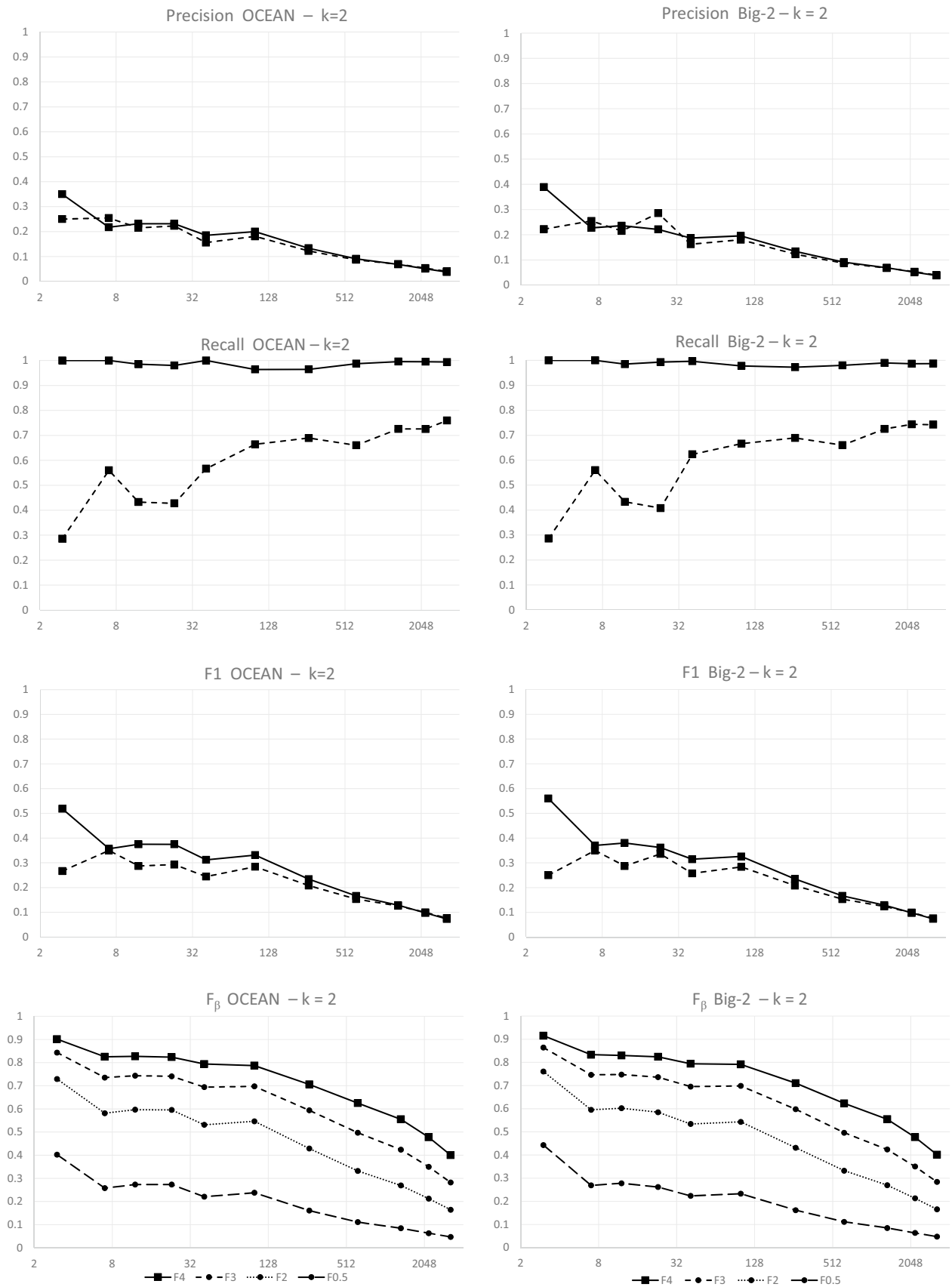

Fig. 9 Continuation of Fig. 8, for $k = 2$

$k = 1$ ($18 \leq$ #Users $\leq 3,266$)

Metric	AUC Big-5	AUC Big-2	Big-2 vs. Big-5
Precision	0.08517	0.10070	18.24%
Recall	0.99416	0.60162	-39.48%
F1	0.15348	0.16731	9.01%

$k = 1$ (#Users ≤ 100)

Metric	AUC Big-5	AUC Big-2	Big-2 vs. Big-5
Precision	0.23779	0.31314	31.69%
Recall	0.98032	0.78378	-20.05%
F1	0.38039	0.44572	17.18%

$k = 2$ ($3 \leq$ #Users $\leq 3,266$)

Metric	AUC Big-5	AUC Big-2	Big-2 vs. Big-5
Precision	0.07488	0.07503	0.21%
Recall	0.99027	0.98519	-0.51%
F1	0.13665	0.13697	0.24%

$k = 2$ (#Users ≤ 100)

Metric	AUC Big-5	AUC Big-2	Big-2 vs. Big-5
Precision	0.18090	0.17947	-0.79%
Recall	0.97201	0.98070	0.89%
F1	0.30346	0.30188	-0.52%

Fig. 10 Comparison of *Area Under the Curve* (AUC) values for the different metrics and values of k. Additionally, we include comparisons for restricted values of numbers of users, corresponding to lower values of the *spread* parameter (and therefore less variability in behavior)

understand and predict people's behavior from social media traces led to an automatization of such processes that relies on the data that is generated by the user through social media platforms and other applications used in mobile devices. In this context, there is no possibility to directly ask questions to the subject, so the values for the different categories and facets that make up the Big-5 taxonomy need to be approximated by analyzing and finding patterns in all kinds of data the users give access to, including texts from posts, geolocation, connections, etc.—as discussed above, bio-inspired computing approaches typically perform quite well in the task of finding such patterns. The first issue here is that a task as complex and qualitative as obtaining insights about an individual's personality is merely quantitatively approximated from a reduced set of data points. Alternatively, in order to reach high precision results, machine learning techniques are used, which generally require a large amount of data to be collected [48, 49]. As we discuss below, from the point of view of explainability and interpretability, as such models usually work as black boxes, it is often difficult—if not impossible—to interpret their results [50, 51].

Finally, though we have already mentioned in Section 1 different works that question issues of privacy for these approaches, there is also the matter of *proportionality*; do we really need to collect the kind and amount of information from a person (even if provided willingly and/or under informed consent) in order to offer them quality services on the Web? The aim of this work is, in part, to show that this may not necessarily be the case, and that accurate approximations that require less intrusive data recollection processes can be used to understand basic human behavior on the Web. The goal of our proposed model is not to capture the whole human personality spectrum, but rather understand those traits that are directly linked to how a person responds to interactions in social media platforms. As our experiments indicate, there is no need to take into account all the traits included in the Big-5 model to address the task of predicting human behavioral patterns in social media—a reduced set of facets affords the same predictive power (or sometimes better) and boasts a greater protection of user privacy. Therefore, for the purposes we focus on, there is no need to build (and to require data for building) such a complete model of a user's personality. Another important observation is that having the capability of obtaining a Big-2 profile does not necessarily give access to a richer set of properties, as afforded by other models.

To illustrate the kind of insights that become available by deriving a Big-5 profile for a user, let's consider the output of the IBM Personality Insights tool over U_1's posts (original post plus ensuing comments)[8] in Fig. 11. Some of the insights obtained include statements like "*You are authority-challenging: you prefer to challenge authority and traditional values to help bring about positive changes*", and "*You are likely to have experience playing music and like historical movies, and unlikely to prefer using credit cards for shopping*". We claim that such conclusions are feasible because the tool has access to the full set of 30 facets (cf. Fig. 2) of the Big-5 model with precise percentiles calculated over a large corpus of users, which can be seen as excessive for simpler applications like the one presented in Section 3.

Finally, another aspect applies to any tool having access to raw data tasked with deriving any personality model is that private information may be unwittingly revealed in the content of social media posts and comments. For instance, U_1 mentions both his/her country of origin (in the original post) and where he/she works (Comment 2). Data privacy should be guaranteed by open-source tools that only make use of the necessary information to carry out their job.

Towards Interpretable and Explainable Models

Other advantages of focusing on a small number of features is that it makes it easier to establish *semantic* connections between the value of a personality feature and the metrics that are actually applied to the subject's data. Though, as discussed in Section 2.3, in our experiments we used a simplified implementation of the Big-2/ROSe model that directly maps a set of Big-5 facets to Big-2 traits, the whole model

[8] This reduced example is for illustrative purposes only—the tool yields only a weak analysis over these 156 words.

Post (U_1) : *I'm very worried that the COVID-19 vaccination in Argentina (my country) is going slowly and that enough people won't be able to get inoculated before the winter.*

> *Comment* 1 (U_2): Don't worry, my friend, I heard in the news that production is ramping up for several of the different vaccines that are out there. If priorities are managed adequately, both the contagion curve and the mortality rates should start to go down soon.

Comment 2 (U_1): Yeah, but still, even if governments get a handle on the supply issues, they still have to deal with the whole anti-vaccine sentiment. I work at City Hall in Buenos Aires, I hear their crazy claims every day.

> *Comment* 3 (U_2): As humans we have always been prone to succumbing to unfounded fears. I think that reaching out to the general public with well-designed non-dismissive campaigns will really help avoid anti-vaccine sentiments.

Comment 4 (U_1): You're so naïve! I think that it should be made mandatory and that people who don't comply be subject to severe fines or even harsher legal penalties.

> *Comment* 5 (U_2): If you think about it, it's natural to be concerned about receiving a vaccine that has been developed in record time, even though preliminary testing has shown it is safe. Don't forget that most people don't know how science works.

Comment 6 (U_1): There's not enough time to be so diplomatic, we're dealing with a global menace that has already changed most people's lives for the worse. Education is at a virtual standstill at all levels, economies are tanking with many jobs being lost, and depression levels are at an all-time high. I'm afraid that we can't stay like this much longer.

> *Comment* 7 (U_2): Don't despair, we are facing something that only occurs every several generations, and has never occurred in this hyper-connected technology-driven world. One thing that has become clear in recent years is that technologies have a way of becoming double-edged swords, so we must address misinformation and malicious activities in general in social media as a way to curb the roots of many social problems.

Fig. 11 Example exchange of views between two users in comments on a social media post

as defined in Section 2 includes a set of three facets for each trait, which we denote F1–F6.

These facets represent a minimal set of characteristics of a user's personality that are directly related to his/her behavior on social media; they encode, to some extent, how users consider or focus on themselves and others, how they relate to others, and the type of sustained behavior they maintain towards others' actions, as well as their own. In a more sophisticated instantiation of our model, we aim for each facet to map to a set of *principles* that describe specific behavior the user needs to show in order to be assigned a value (high/low, or numeric) for the facet. These principles, which could be written in an informal or pseudo-formal language, can later be translated into concrete quantitative metrics to be applied to the available data. Figure 12 provides an overview of the proposed architecture.

Having defined a set of traits and corresponding facets of interests, the most difficult part of such formalization is two-fold: first, we need to be able to establish how to design those principles so that *degrees* to which they hold can be mapped to values for the facets; second, we must specify the concrete set of metrics that need to be applied to the data to determine the degree of satisfaction of each principle, aiming to inspect the least amount of personal data as possible. The complete formalization of the model in these terms is outside the scope of this work, but in the following we show examples of the form that such principles and metrics could have in the setting of a user in a standard social platform, reacting to interactions with other users.

Consider facet *Smoothness* (F5), the facet of the RO trait that describes a person's tendencies toward friendliness, diplomacy, and empathy. High levels of F5 relate to signs of diplomacy and empathy. The following is an example of a set of principles that could guide the valuation of the facet for a particular user; we refer to users U_1 and U_2 in Fig. 11 as examples in each case:

– P_1 – *Use of language*: The use of neutral or positive (negative, vulgar, resp.) vocabulary and tone when posting or interacting with other users, demonstrates signs of diplomacy (antipathy, respectively). User U_1 tends to be negative in his/her tone (for instance, taunting U_2

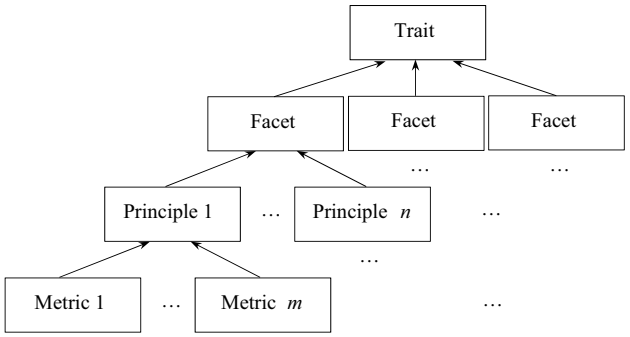

Fig. 12 Structure of proposed interpretable and explainable implementations of the Big-2 model. Values for traits are derived from those of facets, which in turn are obtained by combining values from a set of *principles* that characterize less abstract features. Finally, each principle's value is obtained by combining the results of a set of concrete quantitative *metrics*

by calling him/her naïve, while U_2 is much more diplomatic.

- P_2 – *Contrariness Tendencies*: A high tendency to engage in arguments and debates, usually contradicting other users' opinions and/or using provocative language or tone, is (usually) a sign of low levels of diplomacy and empathy. On the other hand, moderate and measured interventions in conversations and debates, mostly using neutral, positive, and/or conciliatory language and tone, is (usually) a sign of consideration towards others, empathy, and diplomacy. In the example, U_1 is combative (for instance, suggesting severe punishment for people against vaccines) while U_2 is conciliatory (a clear example is his/her suggestion to carry out campaigns that are non-dismissive towards these individuals).
- P_3 – *Positive Social Engagement*: High display of positive content on a user's feed, inviting others to interact and communicate, usually shows signs of friendliness and intentions to reach others in a positive way. Again, we can see in the brief interaction that U_1 is unfriendly, while U_2 shows several signs of empathy ("*Don't worry, my friend, ...*", and "*Don't forget that most people don't know how science works*)".

Principles P_1, P_2, and P_3 aim to capture, in general terms, the connection between the abstract (generally only qualitatively perceived) personality features of a user (e.g., smoothness, diplomacy, empathy) to behavioral features (e.g., engagement in debates, use of specific vocabulary or tone, etc.). Of course, this set of principles is by no means complete and they are not meant to be strict rules; on the contrary, they represent *defeasible associations* that hold in general but there may exist exceptions since human behavior is rarely governed by hard rules. Furthermore, the level of abstraction of such principles is still quite high, and clearly not directly verifiable for a given user. For this purpose, we need to concretize such principles into a set of metrics (then grouped into indicators), that can be directly calculated over the user's available data. As mentioned in Section 1, this is where bio-inspired computing techniques are typically successful when applied over raw data (Step 1 in Fig. 1).

As an example, take principle P_1; one way to deem to which extent P_1 holds for a certain user u is to measure, from all recorded interactions, which percentage exhibits foul, strong, or offensive words (for instance, the outburst of U_1 calling U_2 naïve in Comment 4). Then, if the percentage is over a certain threshold, we can declare user u to have an unfriendly or undiplomatic tendency. Figure 13 shows an extension of this example, identifying the elements for each level of abstraction. There are many sets of principles that can be defined for a given facet, which will in general also depend on the specific set of social platforms that the model will be implemented on, and many concrete criteria to check in the data. The next step in this line of research is the definition of a framework where these three levels of abstraction—(1) facets, (2) principles, and (3) metrics, indicators, and testable criteria—can be formally stipulated. Figures 12 and 13 provide a blueprint of the model; we can think of facets and traits as indicators that combine the results of quantifying several behavioral features via concrete metrics over manifestations of the user's behavior (the available data).

Level of Construction	**Example**
Facets (Indicators) – *Abstract*	*Smoothness* calculated as the average of values obtained with M_1–M_3
Principles (Behavioral Features) – *Less abstract*	P_1: Use of language P_2: Contrariness tendencies P_3: Positive social engagement
Metrics – *Concrete activity-feature connections*	M_1: % of interactions using foul, strong, or offensive words M_2: 1 if user engaged in at least 5 different discussions, 0 otherwise M_3: % of posts with positive sentiment attached

Fig. 13 The three proposed levels used for making the connection between user activity in online platforms and the basic elements of our model. At the highest level we have the Big-2 facets; for each facet, we define a set of *principles*, which characterize specific behavioral features; then, one or more *metrics* are defined for each principle to make a concrete connection with such features and user activity. Finally, metrics are combined into an indicator then used to assign a value to the facet

The advantage of such a framework is that it is capable of providing a clear understanding of *how* the specific data available for a user defines the value for the facets in the Big-2 model. The framework allows to *trace back* predictions to specific behavioral features and concrete values measured over data. For instance, if we predict that a user will react negatively to a certain type of interaction due to the detection of low values of selflessness and smoothness in his/her personality, we can explain *how* this conclusion is reached by using the degrees of satisfaction of the underlying principles, which in turn can be verified by carrying out the associated observations of concrete data. Another benefit of this setup it that is allows for *human-in-the-loop* systems in which—for instance—a surprising result could lead a human analyst to realize that old or otherwise unreliable data is being used. In this case, such data could be discarded and the results recomputed.

Conclusions and Future Work

In this work, we have proposed Big-2/ROSe, a novel, lightweight model for representing the personality of users of online platforms with respect to their activity. After presenting the definition of the traits and facets, we report on the results of a quantitative evaluation showing that a baseline implementation of the new model is capable of performing as well as (and in some cases better than) the Big-5 model when used as a feature in a machine learning classifier trained to make basic predictions of user reactions to Twitter feed content. We then discuss the results of a detailed qualitative evaluation of Big-2/ROSe, highlighting its advantages with respect to privacy, proportionality, interpretability, and explainability.

Ongoing and future work in this line of research and development involves carrying out implementations based on our proposed three-tier framework, and evaluating their performance on real-world data, both in terms of its added value as a feature in machine learning based tools (as done in this paper and our prior work) to address issues of importance like curbing malicious behavior in social media, as well as more abstract evaluations with respect to its capability of modeling different types of users in online platforms.

Acknowledgements We are grateful to V.S. Subrahmanian for providing the Twitter dataset used in our empirical evaluation, and to Constanza Caorsi from IBM Argentina for facilitating access to an academic license for IBM Cloud Personality Insights.

Funding Information This research was funded by Universidad Nacional del Sur (UNS) under grants PGI 24/N046 and PGI 24/ZN34, Secretaría de Investigación Científica y Tecnológica, Facultad de Ciencias Exactas y Naturales, UBA (RESCS-2020-345-E-UBA-REC) by Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) under grant PIP 11220170100871CO, by Agencia Nacional de Promoción Científica y Promoción Tecnológica under grants PICT-2018-0475 (PRH-PIDRI-2014-0007) and PICT-2016-0215 (III-A-Raíces).

Declarations

Conflicts of Interest None of the authors have any conflicts of interest.

Research Involving Human Participants and/or Animals This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent Does not apply, as no studies with human participants were carried out.

References

1. Furr RM. Psychometrics: An introduction. Sage Publications 2017.
2. Allcott H, Gentzkow M. Social media and fake news in the 2016 election. National Bureau of Economic Research: Tech. rep; 2017.
3. Del Vicario M, Zollo F, Caldarelli G, Scala A, Quattrociocchi W. Mapping social dynamics on Facebook: The Brexit debate. Social Networks. 2017;50:6–16.
4. Digman JM. Personality structure: Emergence of the five-factor model. Annual Review of Psychology. 1990;41(1):417–40.
5. Gallo FR, Simari GI, Martinez MV, Falappa MA. Predicting user reactions to Twitter feed content based on personality type and social cues. Future Generation Computer Systems. 2019.
6. Hormes JM, Kearns B, Timko CA. Craving Facebook? Behavioral addiction to online social networking and its association with emotion regulation deficits. Addiction. 2014;109(12):2079–88.
7. Abaeikoupaei N, Al Osman H. A multi-modal stacked ensemble model for bipolar disorder classification. IEEE Transactions on Affective Computing (Early Access). 2020.
8. James TL, Lowry PB, Wallace L, Warkentin M. The effect of belongingness on obsessive-compulsive disorder in the use of online social networks. J Manag Info Sys. 2017;34(2):560–96.
9. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. Journal of Computational Science. 2011;2(1):1–8.
10. Asur S, Huberman BA. Predicting the future with social media. In: Proc. WI. IEEE Comp Soc. 2010;492–499.
11. Galesic M, de Bruin WB, Dumas M, Kapteyn A, Darling J, Meijer E. Asking about social circles improves election predictions. Nature Human Behaviour. 2018;2(3):187.
12. Sanders E, de Gier M, van den Bosch A. Using demographics in predicting election results with Twitter. In: Proc. SOCINFO. Springer. 2016;259–268
13. Kušen E, Strembeck M. Politics, sentiments, and misinformation: an analysis of the Twitter discussion on the 2016 Austrian presidential elections. Onl Soc Netw Med. 2018;5:37–50.
14. Singh P, Dwivedi YK, Kahlon KS, Pathania A, Sawhney RS. Can Twitter analytics predict election outcome? an insight from 2017 punjab assembly elections. Government Information Quarterly. 2020;101444
15. Zhang J, Otomo T, Li L, Nakajima S: Cyberbullying detection on Twitter using multiple textual features. In: 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST). 2019;1–6.

16. Balakrishnan V, Khan S, Fernandez T, Arabnia HR. Cyberbullying detection on Twitter using Big Five and Dark Triad features. Personality and Individual Differences. 2019;141:252–7.
17. Phillips L, Dowling C, Shaffer K, Hodas NO, Volkova S. Using social media to predict the future: A systematic literature review. CoRR abs/1706.06134. 2017;1–55
18. Businessballs: How to be self-aware (course): Personality theories and types. 2021. https://www.businessballs.com/self-awareness/personality-theories-and-types/. Accessed 04-Feb-2021
19. Golbeck J, Robles C, Edmondson M, Turner K. Predicting personality from Twitter. In: Proc. PASSAT@SocialCom.. IEEE. 2011;149-156.
20. Bai S, Hao B, Li A, Yuan S, Gao R, Zhu T. Predicting Big-Five personality traits of microblog users. In: Proc. WI. IEEE Comp Soc. 2013;501–508.
21. Kaushal V, Patwardhan M. Emerging trends in personality identification using online social networks - a literature survey. ACM Trans Knowl Disc Data (TKDD). 2018;12(2):15.
22. Chen J, Haber EM, Kang R, Hsieh G, Mahmud J. Making use of derived personality: The case of social media ad targeting. In: Proc. ICWSM. 2015;51–60.
23. Correa T, Hinsley AW, De Zuniga HG. Who interacts on the Web?: The intersection of users' personality and social media use. Computers in Human Behavior. 2010;26(2):247–53.
24. IBM: IBM Cloud Docs / Personality Insights. https://console.bluemix.net/docs/services/personality-insights/models.html#models. Accessed: 14-May-2020
25. Arnoux PH, Xu A, Boyette N, Mahmud J, Akkiraju R, Sinha V. 25 tweets to know you: A new model to predict personality with social media. In: Proc. ICWSM. 2017.
26. Liu Z, Wang Y, Mahmud J, Akkiraju R, Schoudt J, Xu A, Donovan B. To buy or not to buy? Understanding the role of personality traits in predicting consumer behaviors. In: Proc. SOCINFO. Springer 2016;337–346
27. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. In: Proc. ICWSM. 2013;1–10.
28. Kayiş AR, Satici SA, Yilmaz MF, Şimşek D, Ceyhan E, Bakioğlu F. Big Five-personality trait and internet addiction: A meta-analytic review. Comp Human Behav. 2016;63:35–40.
29. Aidt T, Rauh C. The Big Five personality traits and partisanship in England. Electoral Studies. 2018;54:1–21.
30. Seidman G. The Big 5 and relationship maintenance on Facebook. J Soc Pers Rel. 2019;36(6):1785–806.
31. Sulaiman A, Jaafar NI, Tamjidyamcholo A. Influence of personality traits on Facebook engagement and their effects on socialization behavior and satisfaction with university life. Info Comm Soc. 2018;21(10):1506–21.
32. Mehta Y, Fatehi S, Kazameini A, Stachl C, Cambria E, Eetemadi S. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In: Proc. of ICDM. 2020.
33. Stachl C, Au Q, Schoedel R, Gosling SD, Harari GM, Buschek D, Völkel ST, Schuwerk T, Oldemeier M, Ullmann T, et al. Predicting personality from patterns of behavior collected with smartphones. Proc Nat Acad Sci. 2020;117(30):17680–7.
34. Stachl C, Pargent F, Hilbert S, Harari GM, Schoedel R, Vaid S, Gosling SD, Bühner M. Personality research and assessment in the era of machine learning. Euro J Pers. 2020;34(5):613–31.
35. Li Y, Kazameini A, Mehta Y, Cambria E. Multitask learning for emotion and personality detection. arXiv preprint arXiv:2101.02346. 2021.
36. Basiri ME, Nemati S, Abdar M, Cambria E, Acharya UR. ABCDM: An attention-based bidirectional cnn-rnn deep model for sentiment analysis. Future Generation Computer Systems. 2021;115:279–94. https://doi.org/10.1016/j.future.2020.08.005.
37. Kumar A, Narapareddy VT, Srikanth VA, Malapati A, Neti LBM. Sarcasm detection using multi-head attention based bidirectional lstm. IEEE Access. 2020;8:6388–97.
38. Akhtar MS, Ekbal A, Cambria E. How intense are you? predicting intensities of emotions and sentiments using stacked ensemble. IEEE Computational Intelligence Magazine. 2020;15(1):64–75. https://doi.org/10.1109/MCI.2019.2954667.
39. Mehta Y, Majumder N, Gelbukh AF, Cambria E. Recent trends in deep learning based personality detection. Artif Intell Rev. 2020;53(4):2313–39.
40. Gallo FR, Abad Santos N, Simari GI, Falappa MA. A desiderata for modeling and reasoning with social knowledge. In: Proc. CACIC. 2015
41. Gallo FR, Abad Santos N, Simari GI, Martinez MV, Falappa MA. Belief dynamics in complex social networks. In: Proc. ASAI-JAIIO 45. 2016.
42. Gallo FR, Simari GI, Martinez MV, Falappa MA, Abad Santos N. Reasoning about sentiment and knowledge diffusion in social networks. IEEE Internet Computing. 2017;21(6):8–17.
43. Eysenck HJ. Dimensions of personality. London: Routledge & Kegan-Paul; 1947.
44. Digman JM. Higher-order factors of the Big Five. Journal of personality and social psychology. 1997;73(6):1246.
45. Cambria E. Affective computing and sentiment analysis. IEEE Intell Syst. 2016;31(2):102–7.
46. Hussain A, Cambria E. Semi-supervised learning for big social data analysis. Neurocomputing. 2018;275:1662–73.
47. Recupero DR, Alam M, Buscaldi D, Grezka A, Tavazoee F. Frame-based detection of figurative language in tweets [application notes]. IEEE Comput Intell Mag. 2019;14(4):77–88.
48. Ilmini K, Fernando T. Persons' personality traits recognition using machine learning algorithms and image processing techniques. Adv Comp Sci. 2016;5:40–4.
49. Philip J, Shah D, Nayak S, Patel S, Devashrayee Y. Machine learning for personality analysis based on big five model. In: V.E. Balas, N. Sharma, A. Chakrabarti (eds.) Data Management, Analytics and Innovation. Springer Singapore. 2019;345–355.
50. Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell. 2019;267:1–38.
51. Nott G. Explainable artificial intelligence: Cracking open the black box of AI. Computer World 4. 2017.

Authors and Affiliations

Gerardo I. Simari[1,2,3] · Maria Vanina Martinez[4,5] · Fabio R. Gallo[1,2] · Marcelo A. Falappa[1,2]

Maria Vanina Martinez
mvmartinez@dc.uba.ar

Fabio R. Gallo
fabio.gallo@cs.uns.edu.ar

Marcelo A. Falappa
mfalappa@cs.uns.edu.ar

[1] Department of Computer Science and Engineering, Universidad Nacional del Sur (UNS), Bahia Blanca, Argentina

[2] Institute for Computer Science and Engineering (ICIC UNS–CONICET), Bahia Blanca, Argentina

[3] Arizona State University, Tempe, Arizona, USA

[4] Department of Computer Science, Universidad de Buenos Aires (UBA), Ciudad Autonoma de Buenos Aires, Argentina

[5] Institute for Computer Science Research (ICC UBA–CONICET), Ciudad Autonoma de Buenos Aires, Argentina